DOCUMENT RESUME

ED 250 332 TM 840 547

AUTHOR Cooper, Peter L.
TITLE The Assessment of Writing Ability: A Review of Research.
INSTITUTION Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-84-12; GREB-82-15R
PUB DATE May 84
NOTE 53p.
AVAILABLE FROM Educational Testing Service, Research Publications R-116, Princeton, NJ 08541
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Correlation; Cost Effectiveness; *Essay Tests; *Evaluation Methods; Intellectual Disciplines; *Multiple Choice Tests; Racial Differences; Scoring; Student Evaluation; Test Reliability; Test Validity; *Writing Evaluation

ABSTRACT

Recent information from established testing programs was used to investigate the nature and limitations of essay and multiple-choice tests of writing ability, the statistical relationship of these types of tests, the performance of population subgroups on each, the possible need of different disciplines for different tests of composition skill, and the cost and usefulness of various writing evaluation strategies. The literature indicates that essay tests are often considered more valid than multiple-choice tests. Although essay tests may sample a wider range of composition skills, the variance in essay test scores can reflect such irrelevant factors as speed and fluency under time pressure or even penmanship. Also, essay test scores are typically far less reliable than multiple-choice test scores. Multiple-choice measures tend to overpredict the performance of minority candidates on essay tests. It is not certain whether multiple-choice tests have the same predictive validity for candidates in different academic disciplines, where writing requirements may vary. Still, there appears to be a close relationship between performance on multiple-choice and essay tests of writing ability. The best measures of writing ability have both essay and multiple-choice sections, but this design can be prohibitively expensive. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED250332

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 547

GRE

GRADUATE RECORD EXAMINATIONS

THE ASSESSMENT OF WRITING ABILITY: A REVIEW OF RESEARCH

Peter L. Cooper

GRE Board Research Report GREB No. 82-15R
ETS Research Report 84-12

May 1984

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

The Assessment of Writing Ability: A Review of Research

Peter L. Cooper

GRE Board Research Report GREB No. 82-15R

May 1984

Copyright © 1984 by Educational Testing Service. All rights reserved.

ABSTRACT

The assessment of writing ability has recently recevied much attention from educators, legislators, and measurement experts, especially because the writing of students in all disciplines and at all educational levels seems, on the whole, less proficient than the writing produced by students five or ten years ago. The GRE Research Committee has expressed interest in the psychometric and practical issues that pertain to the assessment of writing ability. This paper presents not a new study but a review of major research in light of GRE Board concerns. Specifically, recent scholarship and information from established programs are used to investigate the nature and limitations of essay and multiple-choice tests of writing ability, the statistical relationship of performances on these types of tests, the performance of population subgroups on each kind of task, the possible need of different disciplines for different tests of composition skill, and the cost and usefulness of various strategies for evaluating writing ability.

The literature indicates that essay tests are often considered more valid than multiple-choice tests as measures of writing ability. Certainly they are favored by English teachers. But although essay tests may sample a wider range of composition skills, the variance in essay test scores can reflect such irrelevant factors as speed and fluency under time pressure or even penmanship. Also, essay test scores are typically far less reliable than multiple-choice test scores. When essay test scores are made more reliable through multiple assessments, or when statistical corrections for unreliability are applied, performance on multiple-choice and essay measures can correlate very highly. The multiple-choice measures, though, tend to overpredict the performance of minority candidates on essay tests. It is not certain whether multiple-choice tests have essentially the same predictive validity for candidates in different academic disciplines, where writing requirements may vary. Still, at all levels of education and ability, there appears to be a close relationship between performance on multiple-choice and essay tests of writing ability. And yet each type of measure contributes unique information to the overall assessment. The best measures of writing ability have both essay and multiple-choice sections, but this design can be prohibitively expensive. Cost cutting alternatives such as an unscored or locally scored writing sample may compromise the quality of the essay assessment. For programs considering an essay writing exercise, a discussion of the cost and uses of different scoring methods is included. The holistic method, although having little instructional value, offers the cheapest and best means of rating essays for the rank ordering and selection of candidates.

Table of Contents

Abstract 1

Introduction ..... [1]

Direct versus Indirect Assessment: The Issues ..... [1]

Sources of Variance in Direct Assessment: Questions of Validity and Reliability 4

- Writer 4
- Topic 4
- Mode 5
- Time limit 5
- Appearance 6
- Examination situation 7
- Rater inconsistencies 7
- Writing context and sampling error 9

Sources of Variance in Indirect Assessment: Questions of Validity and Reliability 11

The Relation of Direct and Indirect Measures 12

- Correlations of scores 12
- Correlated errors 17
- Writing characteristics addressed 17

Group Comparison by Direct and Indirect Assessment 21

Writing Assessment for Different Disciplines 23

Assessment by Direct, Indirect, or Combined Methods 24

- Direct measure 24
- Indirect measure 27
- Combined measures 29

Direct and Indirect Assessment Costs 30

Compromise Measures 32

Scoring Methods for Direct Assessment 34

Descripton of the methods and their uses 34

Advantages and disadvantages of the methods 36

Summary and Conclusion 40

References and Bibliography 42

Introduction

Educators in all disciplines and at all academic levels have become concerned in recent years about the apparent decline in the writing ability of their students, and now legislators share that concern. At present, 35 states have passed or are considering legislation that mandates the statewide testing of students' writing skills (Notes from the National Testing Network in Writing, October 1982, p. 1). As part of a recent research study, faculty members from 190 departments at 34 universities in Canada and the United States completed a questionnaire on the importance of academic writing skills in their fields: business management, psychology, computer science, chemistry, civil engineering, and electrical engineering. In all six areas, writing ability was judged important to success in graduate training. One can assume that it is not merely important but essential in the liberal arts. Even first-year students in programs such as electrical engineering must write laboratory reports and article summaries. Long research papers are commonly assigned to graduates in business, civil engineering, psychology, the life sciences, and of course the humanities and the social sciences. Respondents in the study also agreed that writing ability is even more important to professional than to academic success (Bridgeman & Carlson, 1983).

Understandably, there has been a growing interest, among legislators and educators as well as writing specialists, in the methods used to measure, evaluate, and predict writing skills. Two distinct methods have evolved: "direct assessment" requires the examinee to write an essay or several essays, typically on preselected topics; "indirect assessment" usually requires the examinee to answer multiple-choice items. Hence direct assessment is sometimes referred to as a "production" measure and indirect assessment as a "recognition" measure. Richard Stiggins (1981) observes that "indirect assessment tends to cover highly explicit constructs in which there are definite right and wrong responses (e.g., a particular language construction is either correct or it is not). Direct assessment, on the other hand, tends to measure less tangible skills (e.g., persuasiveness), for which the concept of right and wrong is less relevant" (p. 3).

Direct versus Indirect Assessment: The Issues

Each method has its own advantages and disadvantages, its own special uses and shortcomings, its own operational imperatives. Proponents of the two approaches have conducted a debate--at times almost a war--since the beginning of the century. The battle lines were drawn chiefly over the issues of validity (the extent to which a test measures what it purports to measure--in this case, writing ability) and reliability (the extent to which a test measures whatever it does measure consistently). At first it was simply assumed that one must test writing ability by having examinees write. But during the 1920s and 1930s, educational psychologists began experimenting with indirect measures because essay scorers (also called "readers" or "raters") were shown to be generally inconsistent, or unreliable, in their ratings. The objective tests not only could achieve

extremely high statistical reliabilities but also could be administered and scored economically, thus minimizing the cost of testing a growing number of candidates.

As late as 1966, Orville Palmer wrote: "Sixty years of [College] Board English testing have amply proved that essay tests are neither reliable nor valid, and that, whatever their faults, objective tests do constitute a reliable and valid method of ascertaining student compositional ability. Such a conclusion was very painfully and reluctantly arrived at" (p. 286). And, perhaps, prematurely arrived at. In a landmark study of the same year, Godshalk, Swineford, and Coffman (1966) showed that, under special circumstances, scores on brief essays could be reliable and also valid in making a unique contribution to the prediction of performance on a stable criterion measure of writing ability.

Since then, direct assessment has rapidly gained adherents. One reason is that, while direct and indirect assessments appear statistically to measure very similar skills, as will be discussed later, indirect measures lack "face validity" and credibility among English teachers. That is, multiple-choice tests do not seem to measure writing ability because the examinees do not write. Also, reliance on indirect methods exclusively can entail undesirable side effects. Students may presume that writing is not important, or teachers that writing can be taught -- if it need be taught at all -- through multiple-choice exercises instead of practice in writing itself. One has to reproduce a given kind of behavior in order to practice or perfect it, but not necessarily in order to measure it. Still, that point can easily be missed, and the College Board has committed itself to supporting the teaching of composition by reinstating the essay in the Admissions Testing Program English Composition Test.

The Board's practice is widely followed. Only about 5 percent of all statewide writing assessments rely solely on objective tests; of those remaining, about half combine direct and indirect measures, but the proportion using direct measures alone has increased in recent years and continues to increase. Writing tests for a college-age population are especially likely to require an essay. Along with the popularity of direct assessment has grown the tendency of writing tests, especially in higher education, to focus on "higher-level" skills such as organization, clarity, sense of purpose, and development of ideas rather than on "lower-level" skills such as spelling, mechanics, and usage. Of course, higher-level skills appear naturally to be the province of direct assessment and lower-level skills the humbler domain of indirect assessment; hence the greater face validity and credibility of essay tests for those who teach English.

Earle G. Eley (1955) has argued that "an adequate essay test of writing is valid by definition...since it requires the candidate to perform the actual behavior which is being measured" (p. 11). In direct assessment, examinees spend their time planning, writing, and perhaps revising an essay. But in indirect assessment, examinees spend their time reading items, evaluating options, and selecting responses. Conse-

quently, multiple-choice tests confound the variables by measuring reading as well as writing skills, or rather editorial and error-recognition skills, some insist, in that the examinees produce no writing. Multiple-choice tests, of course, do not require -- or even permit -- any such performance. And they have also been criticized for making "little or no attempt to measure the 'larger elements' of composition, even indirectly" (Braddock, Lloyd-Jones, & Schoer, 1963, p. 42). Their critics generally maintain that objective tests cannot assess originality, creativity, logical coherence, effectiveness of rhetorical strategy, management and flexibility of tone, the ability to generate ideas and supporting examples, the ability to compose for different purposes and audiences, the ability to stick to a subject, or the ability to exercise any other higher-level writing skill -- in short, that objective tests cannot assess anything very important about writing.

Charles Cooper and Lee Odell (1977), leaders in the field of writing research, voice a position popular with the National Council of Teachers of English in saying that such examinations "are not valid measures of writing performance. The only justifiable uses for standardized, norm-referenced tests of editorial skills are for prediction or placement or for the criterion measure in a research study with a narrow 'correctness' hypothesis." But even for placement they are less valid than a "single writing sample quickly scored by trained raters, as in the screening for Subject A classes (remedial writing) at the University of California campuses" (p. viii).

But Cooper and Odell make some unwarranted assumptions: that validity equals face validity, or that what appears to be invalid is in fact invalid, and that one cannot draw satisfactory conclusions by indirect means -- in other words, that one cannot rely on smoke to indicate the presence of fire. Numerous studies show a strong relationship between the results of direct and indirect measures, as will be discussed in greater detail below. Although objective tests do not show whether a student has developed higher-order skills, much evidence suggests the usefulness of a well-crafted multiple-choice test for placement and prediction of performance. Moreover, objective tests may focus more sharply on the particular aspects of writing skill that are at issue. Students will pass or fail the "single writing sample quickly scored" for a variety of reasons -- some for being weak in spelling or mechanics, some for being weak in grammar and usage, some for lacking thesis development and organization, some for not being inventive on unfamiliar and uninspiring topics, some for writing papers that were read late in the day, and so on. An essay examination of the sort referred to by Cooper and Odell may do a fair job in ranking students by the overall merit of their compositions but fail to distinguish between those who are gramatically or idiomatically competent and those who are not.

Sources of Variance in Direct Assessment: Questions of Validity and Reliability

Several sources of score variance reduce the validity and reliability of essay tests.

Writer: Braddock et al. (1963) say that composition exams, "often referred to as measures of writing ability, ...are always measures of writing performance; that is, when one evaluates an example of a student's writing, he cannot be sure that the student is fully using his ability, is writing as well as he can." The authors cite several studies on this point, especially an unpublished Ed.D. dissertation by Gerald L. Kincaid (1953, cited in Braddock et al., 1963) that demonstrates how "the day-to-day writing performance of individuals varies, especially the performance of better writers." If the performance of good writers varies more than that of poor writers, then "variations in the day-to-day writing performance of individual students [do not] 'cancel each other out' when the mean rating of a large group of students is considered" (pp. 6 - 7). In addition to the "writer variable" as a source of error, Braddock et al. identify "the assignment variable, with its four aspects: the topic, the mode of discourse, the time afforded for writing, and the examination situation" (p. 7).

Topic: If no topic is prescribed, it is not psychometrically defensible to compare results. And, too, the variable being measured -- writing ability -- is confounded by other variables, such as the ability to generate a reasonably good topic on the spur of the moment. A test of writing ability is invalid to the extent that the scores reflect anything besides the students' writing ability. But prescribed topics entail their own sorts of difficulties. As Gertrude Conlan (1982) observes, "Between the time an essay question is readied for printing and the time it is administered, any number of things can happen to affect the quality and variety of responses readers must score" (p. 11). For example, developments in the "real world" may make the topic obsolete or emotionally charged. On the other hand, some topics manifest a surprising power to elicit platitudes and generalities; the papers "all sound alike," and readers find it nearly impossible to keep their attention sharp, apply consistent standards, and make sound discriminations. Of course, such circumstances lower the reading and the score reliability of the examination -- that is, the extent to which another reading session would produce the same scores for the same essays or for the same students writing on a new topic.

Score reliability is typically a problem in direct assessment because different topics often require different skills or make different conceptual demands on the candidates. In such cases, comparability of scores across topics and administrations is hard to achieve. As Palmer (1966) remarks, "English teachers hardly need to be told that there exists a great deal of variability in student writing from one theme to another and from one essay to another. The most brilliant students may do well on one essay topic and badly on another. An objective test greatly reduces this inconsistency or variability" (p. 288) by sampling a wider and better

controlled range of behavior with a number of items. Braddock et al. (1963) cite various studies demonstrating that "the topic a person writes on affects the caliber of his writing" (p. 17). By offering only one topic, examiners apparently introduce an invalid source of variance.

For this reason, some researchers have recommended offering a variety of topics if the test population has a broad range of abilities (Wiseman & Wrigley, 1958). Others consider it the lesser evil to provide only one topic per administration. Godshalk et al. (1966) observe, "In the first place, there is no evidence that the average student is able to judge which topic will give him the advantage. In the second place, the variability in topics...would be introducing error at the same time that students might be eliminating error by choosing the topic on which they were most adequately prepared" (pp 13-14). The problem, of course, is exacerbated if the alternate topics are not carefully matched.

Mode: Just as any given writer is not equally prepared to write on all topics, so he or she is not equally equipped to write for all purposes or in all modes of discourse, the most traditional being description, narration, exposition, and argumentation. Variations in mode of discourse "may have more effect than variations in topic on the quality of writing," especially for less able candidates. Even sentence structure and other syntactic features are influenced by mode of discourse (See pages 8, 93, 17 in Braddock et al., 1963). Edward White (1982) says, "We know that assigned mode of discourse affects test score distribution in important ways. We do not know how to develop writing tests that will be fair to students who are more skilled in the modes not usually tested" (p. 17) -- or in the modes not tested by the writing exercise they must perform. Research data presented by Quellmalz, Capell, and Chih-Ping (1982) "cast doubt on the assumption that 'a good writer is a good writer' regardless of the assignment. The implication is that writing for different aims draws on different skill constructs which must therefore be measured and reported separately to avoid erroneous, invalid interpretations of performance, as well as inaccurate decisions based on such performances" (pp. 255-56). Obviously, any given writing exercise will provide an incomplete and partially distorted representation of an examinee's overall writing ability.

Time limit: The time limit of a writing exercise, another aspect of the assignment variable, raises additional measurement problems. William Coffman (1971) observes that systematic and variable errors in measurement result when the examinee has only so many minutes in which to work: "What an examinee can produce in a limited time differs from what he can produce given a longer time [systematic error], and the differences vary from examinee to examinee [variable error]" (p. 276). The question arises, How many minutes are necessary to minimize these unwanted effects?

There is no clear consensus. The short essay has defenders, as will be discussed later, but it also has detractors. Richard Lloyd-Jones

(1977) argues that "a 55-minute test period is still only 55 minutes, so [hour-long] tests are limited to extemporaneous production. That does not demonstrate what a serious person might be able to do in occasions which permit time for reconsideration and revision" (p. 44). Most large-scale assessments offer far less time. The common 20- or 30-minute allotment of time has been judged "ridiculously brief" for a high school or college student who is expected to write anything thoughtful and polished. Braddock et al. (1963) question the validity and reliability of such exercises: "Even if the investigator is primarily interested in nothing but grammar and mechanics, he should afford time for the writers to plan their central ideas, organization, and supporting details; otherwise, their sentence structure and mechanics will be produced under artificial circumstances" (p. 9). These authors recommend 70 to 90 minutes for high school students and 2 hours for college students.

Davis, Scriven, and Thomas (1981) distinguish the longer essay test from the 20- to 30-minute "impromptu writing exercise" in that the latter chiefly measures "students' ability to create and organize material under time pressure" (p. 213). Or merely to transcribe it on paper. Coffman (1971) remarks, "To the extent that speed of handwriting or speed of formulating ideas or selecting words is relevant to the test performance but not to the underlying abilities one is trying to measure, the test scores may include irrelevant factors" (p. 284). The educational community, recalls Conlan, criticized the time limit "when the 20-minute essay was first used in the ECT [English Composition Test],....and the criticism is, if anything, even more vocal today, when writing as a process requiring time for deliberation and revision is being emphasized in every classroom" (Conlan, 3/24/83 memo).

Because time is limited, critics argue, the writing is performed under conditions that approximate those faced by only the most procrastinating and desperate of students or by students taking a classroom examination. Besides being unable to revise carefully, the examinee may be committed to following the first organizational or rhetorical strategy that suggests itself, regardless of whether a better one occurs later. Even apt structural plans may not come to fruition: the student tries to cram half of what there is to say into the last few minutes, and the well-conceived peroration never sees light. Too much time, although less frequently a problem, can have equally unfortunate effects. The student feels the need to keep writing so as not to waste the opportunity and merely recovers old ground, thus damaging the organization of the essay and risking the reader's boredom, if not ire.

Appearance: The ability to write neatly is perhaps even less relevant than the ability to write quickly, but several studies have demonstrated the effect of handwriting quality on essay scores. Lynda Markham (1976) reports that "analysis of variance indicated that the variation in scores explained by handwriting was significant. Papers with better handwriting consistently received higher scores than did those with poor handwriting regardless of [predetermined] quality" (p. 277). McColly (1970) adds,

"There is a tendency for readers to agree in their reactions to appearance. This situation is quite ironic: where the papers that are read are in handwriting, a reliability estimate in the form of a coefficient of interrater agreement will be spuriously high because of the correlation of the appearance factor in the data. At the same time, the validity is actually lowered, because handwriting ability and writing ability are not the same thing" (pp. 153-154).

Examination situation: The performance of a candidate can also reflect the examination situation. This broad aspect of the assignment variable can include the health and mood of the examinee, the conditions in the examination area, and the degree of test anxiety the examinee suffers. It is plausible, but not proved, that these circumstances operate more strongly in direct than indirect assessment. For example, a good writer could "freeze" when confronted with an uncongenial topic, an unfamiliar mode, or a strict time limit, but probably would not draw a blank on an objective test with many items. In view of these considerations, J. K. Hall (1981) and others strongly discourage the "attempt to collect all writing samples in one day" (p. 34). Implicit here is the recognition that the results of essay tests can be very unreliable, or inconsistent, at least when compared to scores on objective tests containing many items.

Rater inconsistencies: Aside from the writer and the assignment variables, the "rater variable" is a major -- and unique -- source of measurement error in direct assessment. "Over and over it has been shown that there can be wide variance between the grades given to the same essay by two different readers, or even by the same reader at different times," observes E. S. Noyes (1963). There is nothing comparable in indirect assessment, except for malfunction by a scanner. Noyes continues, "Efforts to improve the standardization of scoring by giving separate grades to different aspects of the essay -- such as diction, paragraphing, spelling, and so on -- have been attempted without much effect" (p. 8). Moreover, the scoring discrepancies "tend to increase as the essay question permits greater freedom of response" (Coffman, 1971, p. 277).

Even a fresh and experienced reader will vary from one paper to the next in the relative weights that he or she gives to the criteria used for judging the compositions. Often such shifts depend on the types of writing skill in evidence. Readers are much more likely to seek out and emphasize errors in poorly developed essays than in well-developed ones. This notorious "halo effect" can work the other way, too. Lively, interesting, and logically organized papers prompt readers to overlook or minimize errors in spelling, mechanics, usage, and even sentence structure. A reader's judgment of a paper may also be unduly influenced by the quality of the previous paper or papers. A run of good or bad essays will make an average one look worse or better than average.

Naturally, uncontrolled variability will increase as readers become tired and bored. Fatigue causes them "to become severe, lenient, or erratic in their evaluations, or to emphasize grammatical and mechanical features but overlook the subtler aspects of reasoning and organization"

(Braddock et al., 1963, p. 11). The weary reader finds it harder to discount personal feelings or pet peeves. Reader fatigue is an important source of error, but "at what point it begins to be felt is simply not known" (McColly, 1970, p. 151) and can be presumed to vary greatly among readers. It has frequently been shown, however, that the longer a reading period is, or the greater the number of days it covers, the lower the grades tend to become. "Thus the student whose paper is read on the first day has a significant advantage" (Coffman & Kurfman, 1968, p. 105). Moreover, says Coffman (1971), conventional methods of statistical analysis may fail to disclose this kind of measurement error. Estimates of test reliability based on product-moment correlations overlook the change in rating standards insofar as it is uniform across readers "because, in the process of computation, the means and standard deviations of the two sets of scores are equated and only the difference in relative value assigned to the same paper is treated as error. If sets of ratings by a single pair of raters are involved, the overestimation [of reliability] may be excessive" (p. 277).

But interreader disagreement, not just single-reader inconsistency, increases with time. Myers, McConville, and Coffman (1966) studied data from a five-day reading at ETS and found that interrater agreement was significantly lower on the fifth day than on any other day. For each day of the reading the authors computed a "single judgment coefficient" that represented "the average reliability among all readers and across all papers" (p. 44). The coefficient for the last day was .264; the mean for all days -- including the last -- was .406. Such a "loss in vigilance" is understandable, but very troublesome. If readers become less reliable as they near completion of an arduous task, the authors conclude, then "there would be an equivalent drop in reliability regardless of how long the reading period was. This implies immediately, of course, that this problem cannot be handled by simply shortening the reading period by any small amount. It would seem that some external source [that is, beyond a heightened resolve to score accurately] would be needed to bolster the reader morale and effort" (p. 53).

The sad lesson is that if scoring inconsistencies are appreciable in the single reader, they are pronounced in the group -- even for a short reading. Some of them are random -- the cumulative effect of all the individual fluctuations -- and some are systematic, as when one group of readers is invariably more harsh or lenient than another. Hunter Breland and Robert Jones (1982) found that the inexperienced readers in their study often assigned higher scores than did the experienced readers. Again, product-moment correlations tend to mask such differences. Research has also indicated that some raters clearly prefer a nominal style and others a verbal style. (Freedman, 1979, p. 336).

Breland and Jones (1982) remark that even when English teachers "agree on writing standards and criteria..., they do not agree on the degree to which any given criterion should affect the score they assign to a paper. When scoring the same set of papers -- even after careful instruction in which criteria are clearly defined and agreed upon --

teachers assign a range of grades to any given paper" (p. 1). Coffman (1971) adds that "not only do raters differ in their standards, they also differ in the extent to which they distribute grades throughout the score scale." Some readers are simply more reluctant to commit themselves to extreme judgments, and so their ratings cluster near the middle of the scale. Hence "it makes a considerable difference which rater assigns the scores to a paper that is relatively good or relatively poor" (p. 276).

It is impossible to eliminate these many sources of error variance, even with the most exacting set of ratings criteria. Conlan (1980) asserts that "in any reading of essays, some attempt must be made to control for these variables: frequent rest breaks, rules for off-topic papers, a system for handling those papers that are unique or emotionally disturbing to the reader. Despite all efforts to do so, however, all the variables cannot be controlled, and it must be accepted that the...reliability of a reading will never match that of a well designed and carefully controlled system of machine scoring" (p. 26).

Writing context and sampling error: As many have argued, however, the issue of validity is more important than the issue of reliability in the comparison of direct and indirect assessments. It is better to have a somewhat inconsistent measure of the right thing than a very consistent measure of the wrong thing. A direct assessment, say its proponents, measures what it purports to measure -- writing ability; an indirect assessment does not. But the issue is less clear cut than it might appear. In direct assessment, examinees are asked to produce essays that are typically quite different from the writing they ordinarily do. If an indirect assessment can be said to measure mostly editorial skills and error recognition, a brief direct assessment can be said to measure the ability to dash off a first draft, under an artificial time constraint, on a prescribed or hastily invented topic that the writer may have no real interest in discussing with anyone but must discuss with a fictional audience. Since most academic writing, especially at the graduate level, involves editorial proofing and revision, the skills measured by objective tests may compare in importance or validity to those measured by essay tests.

Most defenders of the essay test acknowledge that the validity of the one- or two-essay test is compromised because the exercise is incomplete and artificial. Lloyd-Jones (1982) comments, "The writing situation is fictional at best. Often no audience is suggested, as though a meaningful statement could be removed [or produced and considered apart] from the situation which requires it. The only exigency for the writer is to guess what the scorer wants, but true rhetorical need depends on a well understood social relationship. A good test exercise may implicitly or explicitly create a social need for the purpose of the examination, but the result may be heavily influenced by the writer's ability to enter a fiction" (p. 3).

If the audience is specified, the candidate is writing to two audiences: a fictitious one and a real one that is meant to be ignored -- the raters. The candidate knows little about either but is often supposed to persuade

the former and please the latter. It is very difficult to handle both tasks simultaneously and gracefully, especially if the candidate does not know either audience's tastes, attitudes, degree of familiarity with the subject, prior experience, and so forth. Obviously, communicating genuine concerns to a known audience is a very different enterprise.

This discussion is not intended to urge the superiority of multiple-choice tests -- merely to point out that the validity of direct assessment and the invalidity of indirect assessment cannot simply be presumed because one requires writing and the other does not. For example, Eley (1955) argues that "it would be wrong...to assume that one essay test is as valid as any other" because they all require "the candidate to perform the actual behavior which is being measured" (p. 11). Some examinations will elicit and measure, somewhat unreliably, unrepresentative behavior. Objective tests are not only more reliable but, in some respects, more fair to the students. For instance, it is possible to adjust objective test scores to negate the differences in difficulty across test forms. In fact, programs such as Advanced Placement, New Jersey Basic Skills, and the College-Level Examination Program link objective test scores to essay scores in order to equate the essay scores themselves. There are no sure ways at present of equating essay test scores alone across different tests or administrations, and so students may be unjustly rewarded or penalized for having happened to take a particular examination at a particular time.

Indirect assessment may also afford the examinees more chances to demonstrate their ability. "The good candidate who errs on a few... [multiple-choice items] has plenty of opportunity to redeem himself; a mistake on one item does not affect any other item. In writing a theme, however, the candidate who makes a false start almost inevitably involves his whole theme in difficulties even though he may be, generally speaking, a good writer" (Noyes, Sale, & Stalnaker, 1945, p. 9). As J.W. French (1961) puts it, "The composition test is almost like a one-item test. Some students may happen to enjoy the topic assigned, while others may find it difficult and unstimulating; this results in error" (p. 4). And Coffman (1971), speaking of "the dangers of limited samplings," adds that "the more questions there are, the less likely that a particular student will be too greatly penalized" (p. 289) by the selection.

The inclusion of many items in the objective test benefits the examiner as well as the examinee. As Richard Stiggins (1981) remarks, "Careful construction and selection of [multiple-choice] items can give the examiner very precise control over the specific skills tested," much greater control than direct assessment allows. "Under some circumstances, the use of a [scored] writing sample to judge certain skills leaves the examiner without assurance that those skills will actually be tested.... Less-than-proficient examinees composing essays might simply avoid unfamiliar or difficult sentence constructions. But if their writing contained no obvious errors, an examiner might erroneously conclude that they were competent" in sentence construction. "With indirect assessment, sampling

error is controlled by forcing the examinee to demonstrate mastery or nonmastery of specific elements of writing" (p. 6).

Sources of Variance in Indirect Assessment: Questions of Validity and Reliability

This section is shorter than the preceding one, not because indirect assessment is less problematic, but because fewer sources of variance in indirect assessment have been isolated or treated fully in the literature. For example, just as a given writer may perform better in one mode than another, so a given candidate may perform better on one objective item type than another. The study by Godshalk et al. (1966) shows that some item types, or objective subtests, are generally less valid than others, but more research is needed on how levels of performance and rank orderings of candidates vary with item type. Similarly, not enough is known about whether the examination situation in general and test anxiety in particular are more important as variables in direct or indirect assessment. Also, as noted, the essay writer must try to guess at the expectations of the raters, but to what extent does the candidate taking a multiple-choice test feel impelled to guess what the item writer wants? This task can seem especially ill-defined, and unrelated to writing ability, when the examinee is asked to choose among several versions of a statement that is presented without any rhetorical or substantive context. At best, scores will depend on how well candidates recognize and apply the conventions of standard middle-class usage.

Some sources of score variance in direct assessment are simply not an issue in indirect assessment. Most notably, the appearance and rater variables are negligible in machine-scored objective tests. Daily fluctuations in a candidate's level of performance, analogous to the writer variable, can be accurately estimated and reported as standard error of measurement or as a component of test score reliability. For a well-designed indirect assessment, achieving and documenting reliability is comparatively simple.

Indirect assessment presents fewer reliability problems than does direct assessment, but it also imposes greater measurement limitations. The multiple-choice test, says Lloyd-Jones (1982), is "even less significant as a measure of writing ability" (p. 3) than the conventional writing exercise with all its faults. Indirect assessment affords greater control over sampling error but permits sampling from only a limited area within the domain of writing skills--and some would say a peripheral area. But, as Robert Atwan (1982) remarks, an excellent and valid test need not "duplicate precisely the skill being tested" (p. 14) to allow for very accurate inferences and forecasts. Objective tests can have content, predictive, and construct validity: that is, they can permit students to give a representative performance and enable examiners to sample a controlled (if limited) range of behavior within the domain being tested; they can indicate future course grades and teacher ratings, often better than essay tests do; and they can correlate highly with certain criterion measures, the more highly the more stable the measure.

The Relation of Direct and Indirect Measures

Correlations of scores: The evidence demonstrating the relationship of objective test scores to criterion scores or essay test scores can be found in numerous studies by test publishers and independent researchers alike. It has long been noted that performance on objective tests can predict performance on essay tests with reasonable accuracy (see, for example, Willig, 1926). The following data were obtained under widely different circumstances, but the correlation between direct and indirect measures proves consistent and at least modest, regardless of educational level.

Researcher(s)	Subjects	N	Correlation
Breland, Colon, & Rogosa (1976)	College freshmen	96	.42
Breland & Gaynor (1979)	College freshmen	819	.63
		895	.63
		517	.58
Huntley, Schmeiser, & Stiggins (1979)	College students	50	.43-.67
Godshalk, Swineford, & Coffman (1966)	High school students	646	.46-.75
Hogan & Mishler (1980)	Third graders	140	.68
	Eighth graders	160	.65
Moss, Cole, & Khampalikit (1981)	Fourth graders	84	.20-.68
	Seventh graders	45	.60-.67
	Tenth graders	98	.75-.76

(Stiggins, 1981, p. 1)

Some of the most extensive and important work done on the relation between direct and indirect assessment was performed in 1966 by Godshalk et al. Their study began as an attempt to determine the validity of the objective item types used in the College Board's English Composition Test (ECT). In their effort to develop a criterion measure of writing ability that was "far more reliable than the usual criteria of teachers' ratings or school and college grades" (p. v), they also demonstrated that under certain conditions, several 20-minute essays receiving multiple holistic readings could provide a reliable and valid measure of a student's composition skills. As the authors explain it, "the plan of the study called for

the administration of eight experimental tests and five essay topics [in several modes of discourse] to a sample of secondary school juniors and seniors and for relating scores on the tests to ratings on the essays" (p.6). The eight experimental tests consisted of six multiple-choice item types and two "interlinear exercises," poorly written passages that require the student "to find and correct deficiencies." Of the five essays used as the criterion measure, two allowed 40 minutes for completion and three were to be written in 20 minutes. The examinations were administered to classes with a normal range of abilities. Complete data were gathered for 646 students, who were divided almost evenly between grades 11 and 12.

The authors realized, of course, that they could draw no conclusions about the validities of the indirect measures without first establishing a reliable criterion measure. To minimize the effects of the writer, assignment, and rater variables, they had students write on five topics and in several different modes. Each of a student's five essays received five independent readings. The sum of 25 readings constituted a student's essay total score. The reading reliability of the essay total scores was estimated to be .921. In other words, "if a second group of 25 readers as competent as the first group were chosen and the papers were read again, it might be expected that the two sets of total scores would produce a correlation of approximately .921." The reported score reliability of .841 "is an estimate of the correlation to be expected if the students were to write five more essays on five new topics and if these essays were read by 25 new readers" (pp. 12-13). This latter figure is comparable to the reliability estimate for parallel forms of multiple-choice tests. Applying a correction for attenuation to the data, Coffman (1966) estimated a score reliability of .914 for a five-topic essay test read with perfect reliability (p. 155).

Of course, essay tests cannot be read with perfect reliability. Nor can they often be read five times or include five topics. The reliability estimates decrease along with the number of topics or readings per topic. If each essay receives two independent readings, the score reliabilities are .65 for three topics, .55 for two topics, and .38 for one topic, the most common form of examination (Coffman, 1966). Godshalk et al. (1966) report that "score reliabilites for single topics read once range from .221 to .308. Reading reliabilities for single topics read once range from .361 to .411" (p. 16). "The reliability of essay scores," the study concludes, "is primarily a function of the number of different essays and the number of different readings included.... The increases which can be achieved by adding topics or readers are dramatically greater than those which can be achieved by lengthening the time per topic or developing special procedures for reading" (pp. 39-40).

Having spent the time, money, and energy necessary to obtain reliable essay total scores, Godshalk et al. could then study the correlations of the objective tests and the separate essay tests with this criterion measure. They found that "when objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove to be highly valid" (p. 40). Scores

on the sentence correction and usage multiple-choice subtests had correlations of .705 and .707, respectively, with essay total score (p. 17). These item types have been used in such higher education program tests as the Law School Admission Test (LSAT) and the Graduate Management Admission Test (GMAT). Although the item types are fairly similar, neither section alone achieved a correlation coefficient as high as that obtained by combining scores (.749) (p. 19). Including the verbal score on the SAT also raised correlations with the criterion about .05 on the average.

The single and especially the multiple correlation coefficients are "very high by the usual standards against which validity coefficients are judged" (p. 24). Even the least valid combination of indirect measures had a higher validity coefficient than any that had previously been reported for the objective sections of the ECT, "not... because current forms of the test are more valid than earlier forms but because the criterion is a far better assessment of writing ability than are such measures as school marks or teachers' ratings" (pp. 18-19). Each of the essays, when given two or more readings, made a small but appreciable contribution to the predictive validity of the multiple-choice sections.

In general, though, the separate essays appeared to be less valid measures than the objective subtests (that is, less well correlated with the criterion). The mean intercorrelation among the five separate essay scores was .515. The mean correlation between the five separate essay scores and the sentence correction score was .551. The figure for usage was .553. On the whole, the scores on the essays, each the result of five readings performed under optimal conditions, correlated less well with each other than with scores on the sentence correction or usage sections. In only one of the 10 possible comparisons between two separate essay scores did the essay scores correlate better with each other than with scores on either one of the multiple-choice subtests (these figures and conclusions are derived from data on p. 53 of Godshalk et al., 1966).

Even so, the direct assessment designed as a criterion measure for the Godshalk et al. study is far more valid and reliable than one that can feasibly be administered as a test. Reliabilities are high because "every reader contributed to the total score for each student so that any difference in standards from reader to reader was eliminated." Such is not possible in large-scale readings. The authors observe that "where one reader rated papers from some students while another reader rated those of other students, any differences in standards would contribute to error" (p. 14). Current reading methods reduce, but cannot eliminate, this source of error by having each paper read twice, in some cases by matched pairs of readers, and by having substantial rating discrepancies resolved by an arbiter. Still, each reader contributes to only a fraction of the scores. Moreover, analysis of the variance of essay scores in the Godshalk et al. study indicated that "the ratings assigned varied from topic to topic. When the score is the sum of such ratings," as is rarely the case, "one does not need to be concerned about such differences." The differences can be significant, however, if students are asked to select one or two of

several topics, or if alternate topics appear in different test forms. Then "a student's rating might depend more on which topic he chose [or which test form he took] than on how well he wrote. Differences among the difficulties of the topics would be part of the error." Analysis of the variance also suggested that a student's score on a given essay "might have depended on when in the period his paper happened to be read as well as on how well he wrote" (p. 13).

The Godshalk et al. study, although an acknowledged "classic," is not definitive. First of all, it was based on relatively crude multiple-choice items. The sentence correction items now used in the GMAT program are more sophisticated and, it may well be, more valid. Also, the students were part of an experiment; they had no real incentive to do their best, as they would in an actual admission test, and we can only presume that they performed up to their capacity. As noted, the students were asked to write five essays and take eight objective subtests, including two interlinear exercises. The same readers scored both the exercises and the essays. Although the prescribed sequence of tests "separated the interlinears from the essay exercises to the greatest extent possible," all tests were administered in the fall of 1961 and "all of the reading was done in a single five-day session" (p. 10). The experience of taking the multiple-choice and especially the interlinear subtests could have affected the students' writing. Furthermore, reading the interlinear exercises and the essays in the same session could easily have influenced the standards of raters with prior experience in reading interlinears but no prior experience "in reading brief extemporaneous essays produced by average eleventh- and twelfth-grade students" (p.10). That is, readers could have been prone to score essays according to how well the students handled or avoided problems of the sort addressed in the interlinears. Perhaps the administration and reading schedules inflated the correlation between direct and indirect measures.

More recent studies, though, have confirmed the high correlations observed by Godshalk et al. Hunter Breland (1977a) reported on the use of the multiple-choice Test of Standard Written English (TSWE) for college English placement in four institutions. This test, comprising mostly usage and sentence correction items, is designed not to distinguish among students at the upper end of the score scale but to assist in decisions about students who may need remedial instruction. Breland found a strong relationship between performances on the TSWE and on essay tests of writing ability. Also, the TSWE predicted writing performance during the freshman year at least as well as did a brief essay test given at the beginning of the year; 90 percent of the examinees in the top score range on the TSWE earned an "A" or "B" in their English course.

Breland and Judith L. Gaynor (1979) sampled over 2,000 students in collecting data for a longitudinal study. Students wrote a 20-minute essay at the beginning, middle, and end of their freshman year. Each essay was on a different topic, and each was scored independently by two experienced readers using a 6-point holistic scale. Students also responded

to different forms of the TSWE on three separate occasions. Analysis of the data shows that each of the three TSWE examinations correlated better with the essay scores, considered singly or in various combinations, than the essay scores correlated with each other.

Correlations of Direct and Indirect Measures of Writing Ability

Variable Number and Variable Name	Mean[a]	S.D.[a]	Variable Number 1	2	3	4	5	6	7	8	9	10	11	12
1. TSWE Pretest	43.08	11.17		819	836	804	493	458	836	752	280	266	286	261
2. Essay Pretest	6.72	2.16	.63		895	942	331	315	717	904	307	310	286	289
3. TSWE Posttest (Fall)	46.03	10.43	.83	.63		926	333	311	836	865	333	292	286	288
4. Essay Posttest (Fall)	7.08	2.23	.58	.52	.56		328	315	747	942	306	315	263	210
5. TSWE Posttest (Spring)	47.82	9.97	.84	.57	.84	.52		517	286	323	[illegible]	298	286	294
6. Essay Posttest (Spring)	7.02	2.44	.58	.51	.62	.50	.58		266	310	297	315	256	310
7. TSWE Total (1 + 3)	89.02	20.76	---[b]	.66	---	.60	.89	.64		697	286	248	286	244
8. Essay Total (2 + 4)	13.58	3.83	.69	---	.68	---	.63	.58	.72		301	310	258	310
9. TSWE Total (3 + 5)	95.60	18.84	.87	.63	---	.60	---	.64	---	.70		278	286	274
10. Essay Total (4 + 6)	14.26	4.13	.69	.60	.71	---	.65	---	.72	---	.72		238	210
11. TSWE Total (1 + 3 + 5)	138.48	29.09	---	.65	---	.62	---	.64	---	.72	---	.73		234
12. Essay Total (2 + 4 + 6)	21.18	5.62	.72	---	.74	---	.69	---	.76	---	.75	---	.76	

Note: Correlations are below the diagonal, N's above. Recomputation for minimum N revealed no important differences from the correlations shown.

[a]Based on maximum N.

[b]Correlations are not shown because part scores are included in total score.

(Breland & Gaynor, 1979, p. 123)

For example, the essay posttest (spring) had correlations of .58 with the TSWE pretest, .62 with the TSWE posttest (fall), and .58 with the TSWE posttest (spring), but correlations of only .51 with the essay pretest and .50 with the essay posttest (fall).

The indirect measure had a predictive advantage over the direct measure, even when a direct measure was the criterion predicted (see, however, discussion of assessment by indirect measures beginning on page 27). Correlations between direct and indirect measures were substantially higher when direct assessments were totalled but not when indirect assessments were totalled. This fact suggests that the correlations between direct and indirect measures, although substantial, are limited chiefly by the reliability of the direct measure. When that reliability is increased with multiple assessments or readings, as was the case in the study by

Godshalk et al. (1966), the correlations can increase dramatically. Breland and Gaynor (1979) wrote that "correcting the observed correlation between direct and indirect composites... (.76) would thus yield an estimated true score correlation of .90 between direct and indirect measures" (p. 126).

Correlated errors: The true score correlations could be higher still. Breland and Gaynor reported "an average score reliability of .51 for a single 20-minute essay topic with two readings" (p. 126), a substantially higher figure than the .38 reported by Coffman (1966). But in 1980, Werts, Breland, Grandy, and Rock argued that the figure of .51 might have been inflated by correlated errors of measurement -- systematic influences of such irrelevant factors as neatness and penmanship (as discussed previously). According to Werts et al., "about 21.2% of a test-retest correlation between any two essay ratings can be ascribed to correlated errors" (p. 27), and they concluded that a more accurate estimate of the reliability reported by Breland and Gaynor is between .35 and .45. Of course, the lower reliability would further attenuate the correlation between direct and indirect measures and require a greater correction. "Even higher estimates [than .90] of the true correlation between direct and indirect assessments are obtained when the effects of correlated errors of measurement are considered" (Breland & Gaynor, 1979, p. 126).

Writing characteristics addressed: These extremely high estimates raise the question of whether the tests are measuring essentially the same skills, or skills that -- if different -- appear together so consistently as to allow an accurate measure of one type to serve as a measure of the other type. Are there, then, real differences between what essay tests and multiple-choice tests evaluate? In seeking to answer this question, Richard Thompson (1976) found that the best predictors of holistic scores on 45 student papers were ratings on three criteria not amenable to multiple-choice testing: "unsupported statement," "independent judgment error," and "lack of unity" (p.1). In 1982, Breland and Jones attempted to determine what aspects of writing skill contribute most to the grades given brief impromptu essays. The authors drew a random sample of 806 20-minute essays out of the 85,000 written for the English Composition Test (ECT) administered by the College Board in December 1979. The essays were read holistically and scored as part of the ECT. Then they received a special holistic reading (the "PWS reading") on September 27, 1980. Twenty college English professors from different parts of the country not only scored the papers but annotated them to indicate the presence or absence of important features and also completed a set of Essay Evaluation Forms by marking each essay as being very strong, strong, weak, or very weak on 20 separate characteristics.

Each characteristic was classified as being either a discourse characteristic, a syntactic characteristic, or a lexical characteristic.

Discourse Characteristics	Syntactic Characteristics	Lexical Characteristics
1. Statement of thesis	10. Pronoun usage	16. Level of diction
2. Overall organization	11. Subject-verb agreement	17. Range of vocabulary
3. Rhetorical strategy	12. Parallel structure	18. Precision of diction
4. Noteworthy ideas	13. Idiomatic usage	19. Figurative language
5. Supporting material	14. Punctuation	20. Spelling
6. Tone and attitude	15. Use of modifiers	
7. Paragraphing and transition		
8. Sentence variety		
9. Sentence logic		

(Breland & Jones, 1982, p. 6)

As the authors define them, "those characteristics grouped as 'discourse' are seen as features of the composition as a whole or of a prose piece at least as long as the conventional paragraph. 'Syntactic' characteristics are those that attach to the sentence, clause, or phrase; 'lexical' characteristics are those of the word or word unit" (p. 6).

The PWS reading scores had a correlation of .58 with the ECT scores on the very same essays. The correlation between a scoring and rescoring of a multiple-choice test would be nearly perfect. "Of course," observe Breland and Jones, "the correlation between ECT holistic and PWS holistic represents an estimate of the [reading, not score,] reliability of each. It is an upper-bound estimate, however, because the same stimulus and response materials were judged in both instances -- rather than parallel forms" (p. 13). On the other hand, the two groups of readers were trained at different times and for somewhat different tasks, a circumstance that could lower reliability. The ECT objective score also had a correlation of .58 with the ECT holistic score. If the reliability of the latter is .58, "correction for attenuation would increase the correlation with the ECT objective score... to .76. Lower estimates of ECT holistic score reliability would attenuate the correlations even more" (p. 13).

Despite the high estimated true correlation, the study does support Thompson's (1976) finding that essay scores depend primarily on higher-order skills or discourse characteristics that multiple-choice examinations cannot measure directly. "Eight of the 20 PWS characteristics proved to contribute significantly to the prediction of the ECT holistic score," and five of these were discourse characteristics.

Multiple Prediction of ECT Holistic Score from Significant Contributors

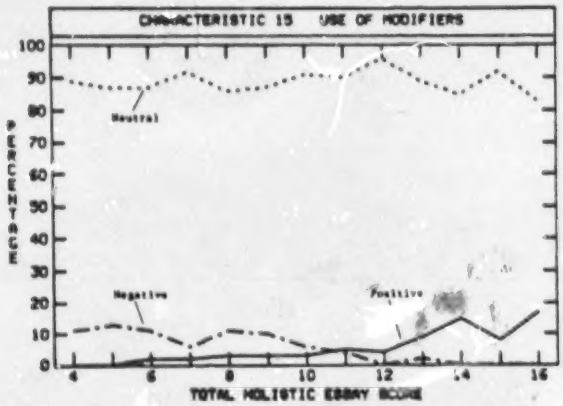
Variables (in order entered, within set)	*r*	*R (cumulative)*	*b*	*β*
Using Significant PWS Characteristics				
2. Overall organization	.52	.52	.15	.20
4. Noteworthy ideas	.48	.57	.09	.12
20. Spelling	.29	.58	.13	.10
7. Paragraphing and transition	.42	.59	.10	.11
12. Parallel structure	.31	.60	.14	.08
5. Supporting material	.47	.61	.08	.12
9. Sentence logic	.39	.61	.07	.07
16. Level of diction	.40	.61	.08	.06
Using ECT Objective and Significant PWS Characteristics				
40. ECT objective score (raw)	.58	.58	.04	.39
2. Overall organization	.52	.67	.14	.18
4. Noteworthy ideas	.48	.68	.10	.14
20. Spelling	.29	.69	.12	.09
7. Paragraphing and transition	.42	.70	.09	.10
12. Parallel structure	.31	.70	.09	.06
9. Sentence logic	.39	.70	.04	.04

"Interestingly, ... the multiple correlation of .61 attained is not substantially higher than that possible using [the five significant] discourse characteristics alone (.59)" (Breland & Jones, 1982, p. 15).

On the other hand, the figure for ECT objective raw score alone is .58 before correction for attenuation, higher than that produced by any combination of two discourse characteristics. When direct and indirect assessments are combined, the multiple correlation jumps to .70, suggesting that for this population the instruments tap closely related but distinct skills. Each measures some things that the other may miss: objective tests cannot evaluate discourse characteristics, and essay tests may overlook important sentence-level indicators of writing proficiency. For example, the PWS Essay Evaluation Forms showed the least variance among essays in characteristic 15, use of modifiers (Breland & Jones, 1982, p. 42). Also, this characteristic had virtually no predictive utility and had the lowest set of intercorrelations with the other writing characteristics (p. 45), largely because the low variance severely attenuated the correlations. But in multiple-choice tests such as sentence correction, use of modifier problems often produce very difficult and discriminating items that distinguish among candidates of high ability.

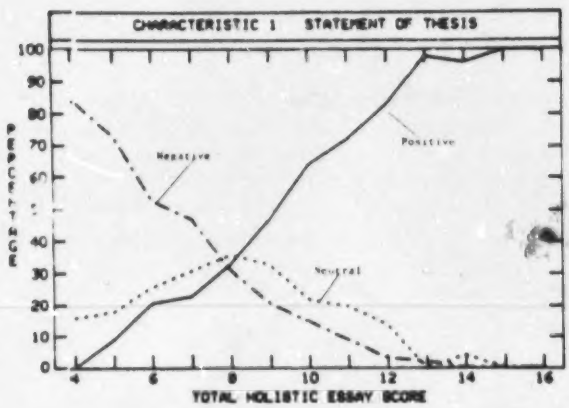
A possible explanation for this seeming contradiction is that essay tests allow students to avoid what they cannot do. Less able writers will usually shirk sophisticated constructions, opting instead for repetitious subject-verb-complement sentences. Modifiers generally dangle as part of a more complex structure. In fact, the writer who rarely employs subordinate clauses, verb phrases, absolute constructions, or similarly advanced syntactic units will probably misuse modifiers very little. Strong candidates handle modifiers adequately, but even they may not be very adventurous when writing hurriedly and without time to revise or

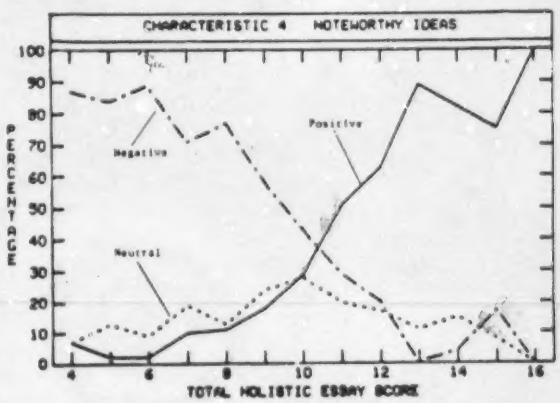
polish. Then their essays will not stand out for arresting use of modifiers. Consequently, the impromptu essay writers appear to use modifiers the same way--neutrally.



(Breland & Jones, 1982, p. 63)

This chart shows that about 90 percent of the essays at every score level, from lowest to highest, were rated neutral on "use of modifiers." Consider the charts for "statement of thesis" and "noteworthy ideas":





(Breland & Jones, 1982, p. 61)

A great many composition experts contend, however, that the use of modifiers is an essential writing skill that reveals much about a student's overall proficiency. Virginia Tufte and Garrett Stewart (1971) deem it "probably the most crucial topic in the whole study of grammar as style" (p. 14). And Francis Christensen (1968), who builds his entire "New Rhetoric" program around the use of modifiers, quotes John Erskine as saying that "the modifier is the essential part of any sentence" (p. 1). Facility with modifiers allows the writer to vary sentence structures -- to be flexible and inventive as well as grammatical.

Also, multiple-choice item types that address modification and other sentence-level problems often show a strong relation with direct measures. Recall that Godshalk et al. (1966) reported a correlation of .749 between the essay total scores and the sums of raw scores on the sentence correction and usage subtests, "a finding that strongly suggests that the characteristics of single sentences play an important part in the overall impression made on the reader" (p. 19). Perhaps differences in sentence level skills influenced the PWS readers more than they realized and reported. Or perhaps the conclusion suggested by the PWS study reflects a shift in standards for judging writing, one brought about by new scoring techniques and theories on teaching composition.

Group Comparisons by Direct and Indirect Assessment

Despite any such shifts, research past and present amply documents the predictive utility of indirect measures of writing ability. Recently, investigators have begun asking whether these measures are equally useful for different population groups, at different ability levels, or in different disciplines. Breland (1977b) noted that the TSWE appeared to predict the freshman-year writing performance of male, female, minority, and nonminority students at least as well as did high school English grades, high school rank in class, or writing samples taken at the beginning of the freshman year. Because of the relatively small number of minority students in the sample, however, all minorities were grouped together for analysis, and no analysis within subgroups was performed.

A 1981 study by Breland and Griswold does not have that limitation and allows for more detailed inferences. The authors reported that "when the nonessay components of the EPT [English Placement Test] were used to predict scores on the EPT-Essay,... tests of coincidence of the regression lines showed significant differences ($p<.01$)"; these differences suggested that "the white regression overpredicts for blacks throughout the range of EPT scores. A similar pattern is evident for prediction of Hispanic scores" (p.8). The essay-writing performance of men was also overpredicted and that of women underpredicted by the multiple-choice measures, which included TSWE and SAT verbal scores as well as EPT nonessay scores. In other words black, Hispanic, and male students at a given score level tended to write less well than the average student at that score level; white students and women especially tended to write better. "Those scoring low on a predictor but writing above-average essays are false

negatives, and those scoring high on a predictor but writing below-average essays are false positives....The highest proportions of false negatives predicted by TSWE are for women and whites, and the lowest proportions for men, blacks, and Hispanics" (p. 11). Data on false positives were not presented. These findings, the authors remark, confirm those of previous studies.

Breland and Jones (1982) had a population sample that allowed them to make additional group comparisons. Their 806 cases consisted of 202 blacks (English best language), 202 Hispanics-Y (English best language), 200 Hispanics-N (English not best language), and 202 whites (English best language). The authors reached a familiar conclusion: "Multiple-choice scores predict essay scores similarly for all groups sampled, with the exception that lower-performing groups tend to be overestimated by multiple-choice scores" (p. 28). Again, "analyses showed that women consistently wrote more superior essays than would be predicted by the multiple-choice measures and that men and minority groups wrote fewer superior essays than would be predicted" (p. 21).

The data also permitted group comparisons within score levels.

Essay Writing Performance by Group and SAT-Verbal Score Level

SAT-V Score Range	*Group*					
	Total	*Men*	*Women*	*Black*	*Hispanic*	*White*
	Frequencies Scoring in Four SAT-V Score Ranges					
500+	37,716	18,714	18,990	786	398	30,541
400-499	24,659	11,650	13,004	896	427	18,898
300-399	9,457	4,319	5,135	617	305	6,269
below 300	1,534	684	848	164	107	602
	Percentages Writing Superior[a] Essays					
500+	55.9	51.2*	60.5*	42.9*	47.7*	56.4
400-499	26.9	21.4*	31.9*	21.9*	20.6*	27.8
300-399	10.9	7.3*	13.9*	8.3*	7.2*	12.3*
below 300	2.4	1.8	2.9	1.2	1.9	4.0

a. A "superior" essay was defined as one receiving an ECT holistic score of 6 or above.

*Statistically significant ($p < .05$) from expected percentages. Direction of effect is determined by comparing percentage with that in "Total" column.

Essay Writing Performance by Group and TSWE Score Level

TSWE Score Range	*Group*					
	Total	*Men*	*Women*	*Black*	*Hispanic*	*White*
	Frequencies Scoring in Four TSWE Score Ranges					
50+	40,692	18,536	22,145	811	412	33,143
40-49	21,905	11,066	10,834	855	419	16,474
30-39	8,578	4,560	4,014	573	259	5,709
below 30	2,191	1,205	984	224	147	982
	Percentages Writing Superior[a] Essays					
50+	54.0	50.6*	56.9*	42.9*	47.3*	54.5
40-49	26.3	23.3*	29.4*	20.4*	20.0*	27.1
30-39	11.2	9.1*	13.5*	10.3	7.7*	12.2
below 30	3.4	3.4	3.4	2.2	2.0	4.5

a. A "superior" essay was defined as one receiving an ECT holistic score of 6 or above.

*Significantly different ($p < .05$) from expected percentage.

(Breland & Jones, 1982, p. 22)

For whites, underprediction decreases somewhat with increasing ability as measured by the SAT verbal and TSWE score levels: the difference between the percentage of whites and of the total group writing superior essays gets smaller as the score levels get higher. For women, underprediction is greatest around the middle of the objective score scale. For Hispanics and especially for blacks, overprediction increases with ability level.

Such was not so clearly the case in the Breland and Griswold (1981) study, but there the sizes of the groups varied greatly. Also, the portions that were sampled may not have been typical of the subgroups: only 64 percent of the students tested reported ethnic identity.

In the Breland and Jones (1982) study, "multiple regression analyses suggested that the influences of specific elements on holistic scores of essays vary across groups. When the groups were compared also with respect to the frequency with which [the 20 PWS writing] characteristics were perceived positively or negatively, a slightly different picture of relative importance resulted" (p. 21). In addition, groups differed in terms of relative strengths and weaknesses. For example, the three largest percentages of negative ratings within the sample of white essays were for precision of diction, rhetorical strategy, and paragraphing and transition. The three largest percentages of negative ratings within the sample of black essays were for noteworthy ideas (22 percent positive, 20 percent neutral, 57 percent negative), supporting material (32 percent positive, 17 percent, neutral, 51 percent negative), and overall organization (30 percent positive, 19 percent neutral, 50 percent negative)--the three characteristics most highly correlated with and predictive of the combined holistic scores from the ECT and PWS readings (pp. 20, 45). Each is much more highly correlated with the total holistic score than with the ECT objective or TSWE score. In fact, the differences between correlations with total holistic and with objective scores are greater for these three characteristics than for any of the others (p. 45). It is not surprising, then, that the indirect measures overpredict black performance on the direct measure.

Writing Assessment for Different Disciplines

The Breland and Jones (1982) study shows that "different writing problems occur with different frequencies [and degrees of influence] for different groups" (p. 27), but it does not indicate how the relative importance of any given characteristic might change with writing task or type of discourse. The question is important because different academic audiences may vary widely in the types of writing they demand and hence in the standards by which they judge student writing (Odell, 1981). Indirect assessments tend to measure basic skills that are deemed essential to good writing in any academic discipline. However, disciplines could disagree on how much weight to give the less tangible and more subjective higher-order qualities of writing that are directly revealed only through composition exercises. If the observed correlation between performance on direct and indirect measures holds steady throughout the range of ability levels when writing exercises reflect the needs of different disciplines, then disciplines requiring different amounts of writing or different levels of writing proficiency would not necessarily need different types of measures to assess the writing ability of their prospective students. On the other hand, if the correlation varies among disciplines when the writing exercises reflect the special needs of each, then different types of measures may be advisable.

Little work has been completed on how the relation between direct and indirect measures varies with writing task. However, Sybil Carlson and Brent Bridgeman (1983) conducted research on the English skills required of foreign students in diverse academic fields. Such information, they maintain, must be gathered before there can be "a meaningful validation study of TOEFL," the multiple-choice Test of English as a Foreign Language. The faculty members they surveyed generally reported relying more on discourse level characteristics than on word or sentence level characteristics in evaluating their students' writing. But when departments were asked to comment on the relevance of 10 distinct types of writing sample topics, the different disciplines favored different types of topics and, by implication, placed different values on various composition skills. For example, engineering and science departments preferred topics that asked students to describe and interpret a graph or chart, whereas topics that asked students to compare and contrast, or to argue a particular position to a designated audience, were popular with MBA and psychology departments as well as undergraduate English programs.

Bridgeman and Carlson (1983) concluded from the results of their research questionnaire: "Although some important common elements among different departments were reported, the survey data distinctly indicate that different disciplines do not uniformly agree on the writing task demands and on a single preferred mode of discourse for evaluating entering undergraduate and graduate students. The extent to which essays written in different discourse modes produce different rank orderings of students remains to be seen. Furthermore, if significant differences in rank ordering are observed, the relationship of these orderings to TOEFL scores, both within and across disciplines and language groups, is yet to be determined" (p. 56). This statement has clear implications for the GRE program. A single measure, direct or indirect, will not serve all disciplines equally well if the rank ordering of students in a direct assessment can be significantly altered by varying the topics or discourse modes to reflect the special needs of each discipline.

Assessment by Direct, Indirect, or Combined Methods

All the issues raised thus far regarding direct and indirect measures bear on one overriding question: What is the best way to assess writing ability--through essay tests, multiple-choice tests, or some combination of methods? The answer to this question involves psychometric, pedagogical, and practical concerns.

Direct measure: The use of a direct measure as the sole test of writing ability may appeal to English teachers, but it may also prove inadequate, especially in a large testing program. Essay examinations, for one thing, simply are not reliable enough to serve alone for making crucial decisions about individuals. A population like that taking the GRE General Test could pose even bigger reliability problems. Coffman (1971) writes, "The level [of reliability] will tend to be lower if the examination is administered to a large group with heterogeneous backgrounds rather than to a

small group whose background of instruction is homogeneous. It will tend to be lower if many different raters have to be trained to apply a common standard of evaluation in reading the papers. It will tend to be lower for less exact and more discursive subject matters," especially for a "discipline-free" or "general composition topic" (pp. 278-279). And as Coffman observes, attaining a high reading reliability through multiple readings "does not insure that score reliability will also be high.... It does not do much good to achieve high rating reliability (usually at considerable expense), for example, if the sample is so inadequate that the test [score] reliability remains low." He reminds us that the rating reliability of a single multiple-choice item "is likely to be in excess of .99, yet nobody would confuse this coefficient with the reliability of a test made up of two or three such items" (pp. 281, 297).

Coffman (1966) also remarks that "placing undue emphasis on reliability when evaluating essay tests...may lead to unwarranted conclusions. If essay tests are to be compared with objective tests, it is the correlations of each with a common reliable criterion measure, not their respective reliability coefficients, which should be compared" (p. 156). But by this standard too, essay tests used alone are found wanting. As Godshalk et al. (1966) report, "For the essays, the first-order validities do not approach the range of validities for other [i.e., the multiple-choice] types of questions until the score is based on three readings" (p. 41). In a field trial follow-up to the original study, the authors determined that even when essays received four readings, the average correlation of an essay's total score with the criterion measure was substantially lower than that of the sentence correction or usage score (p. 34).

These findings confirm "the common opinion that a single topic is inadequate" (McColly, 1970, p. 152) on grounds of validity and reliability for assessing an individual's writing ability. Kincaid established as early as 1953 that "a single paper written by a student on a given topic at a particular time cannot be considered as a valid basis for evaluating his achievement," especially if the student is above average (cited in Braddock et al., 1963, p. 92). Lee Odell (1981) argues, "The ability to do one sort of writing task may not imply equal ability with other kinds of tasks.... We need to evaluate several different kinds of writing performance," unless only one kind (for example, the persuasive essay) is of interest. Moreover, "we must have at least two samples of the student's writing for each kind of writing" (pp. 115, 118). And Paul Diederich has maintained that two themes are "totally inadequate" (cited in Braddock et al., 1963, p. 7n). They certainly appear so unless each is given numerous readings.

Research by Coffman (1966) has compared the importance of multiple topics to that of multiple readings.

Estimates of Validity (r_{ie}), Score Reliability (r_{ii}), and Reading Reliability (r_{aa}) for Sets of Short Essays Read Holistically*

Number of Readings per Topic		Number of Topics				
		1	2	3	4	5
1	r_{ie}	.462	.581	.648	.691	.721
	r_{ii}	.263	.416	.517	.588	.641
	r_{aa}	.386	.513	.598	.657	.701
2	r_{ie}	.555	.668	.724	.758	.782
	r_{ii}	.380	.550	.647	.709	.753
	r_{aa}	.557	.679	.748	.792	.824
3	r_{ie}	.601	.707	.757	.786	.805
	r_{ii}	.445	.616	.706	.762	.800
	r_{aa}	.653	.760	.816	.851	.876
4	r_{ie}	.628	.728	.777	.801	.818
	r_{ii}	.487	.655	.740	.792	.826
	r_{aa}	.715	.808	.855	.885	.904
5	r_{ie}	.648	.744	.787	.811	.826
	r_{ii}	.517	.682	.763	.811	**.841
	r_{aa}	.759	.842	.882	.906	**.921
		*****				
∞	r_{ie}	.743	.810	.838	.852	.861
	r_{ii}	.681	.810	.865	.895	.914
	r_{aa}	1.000	1.00	1.000	1.000	1.000

* Estimates are based on data from 646 secondary school juniors and seniors who each wrote five short essays. The hypothetical criterion is a set of four additional essays each read four times.

**These coefficients are based on empirical data.

(Coffman, 1966, p. 154)

The value of reading reliability (r_{aa}), naturally, depends more on the number of readings per topic than on the number of topics read. For example, r_{aa} =.701 for five topics read once each and .759 for one topic read five times. But values for validity (r_{1c}) and score reliability (r_{11}) depend more on the number of topics read than on the number of readings per topic. Since the validity of an examinee's score is clearly more important than its reading reliability, it would seem necessary for a valid direct assessment to require several essays from each student. Note, however, that values for all figures tend to be highest when R (number of readings per topic) times T (number of topics) is maximum for any number n where n=R+T. For example, two topics read twice each yield higher figures than either one topic read three times or three topics read once each. Coffman (1966) considers the validity coefficients reported by Godshalk et al. (1966) for the one-hour objective composite of subtests and concludes, "In order to obtain validity coefficients of comparable magnitude using only essay tests, it would be necessary to assign at least two topics to each student and have each read by five different readers or to assign three topics and have each read by three different readers" (p. 156). Large-scale assessments rarely find it possible to give more than one topic with two independent readings.

Indirect measure: The essay examination, then, appears unworkable as the sole measure of writing ability in a program that seeks to rank order candidates for selection. Scores based on two readings do not allow for reliable close discriminations but can be useful when considered along with other information. Could indirect assessment alone be adequate for the purpose of ordering and selecting candidates? Edward Noyes declared in 1963 that well-crafted objective English items "measure how well a student can distinguish between correct and incorrect expression...and how sensitive he is to appropriate or inappropriate usage"; objective English tests are "based on the assumption that these skills are so closely associated with actual writing ability that direct measurement of them will provide a reasonably accurate, if indirect, measurement of writing ability" (p. 7). As discussed, numerous studies bear out this assumption. In particular, Godshalk et al. (1966) reported, "We have demonstrated that the [objective] one-hour English Composition Test does an amazingly effective job of ordering students in the same way as a trained group of readers would after reading a sizable sample of their actual writing." Various multiple-choice item types proved "remarkably effective for predicting scores on a reliable criterion of writing ability" (pp. 21, 29). The authors added, "one hypothesis of this study was that the multiple-choice questions were discriminating primarily at the lower levels of skill.... No convincing support for this hypothesis was found" (p. 42). The objective sections worked very well alone and even better when combined with SAT or PSAT verbal scores.

From a psychometric point of view, it does appear that indirect assessment alone can afford a satisfactory measure of writing skills for ranking and selection purposes. Many argue persuasively that it is inadequate by itself for such purposes as charting the development of individual student writers, gauging the impact of a new writing program,

gaining a full picture of a given student's strengths and limitations in various writing situations, or placing a student in one of several English classes. During the period 1977-1982, ETS's Programs for the Assessment of Writing contracted with a number of colleges to develop direct assessments. These colleges were not satisfied with their record of placement decisions based solely on TSWE scores. Richard Stiggins (1981) adds that even for selection, direct assessment is to be preferred "when writing proficiency is the sole or primary selection criterion," as for scholarships to a college writing program; indirect measures "are acceptable whenever writing proficiency is one of the many selection criteria" (p. 11).

Still, the use of multiple-choice tests alone for any writing assessment purpose will elicit strong objections. Some, such as charges that the tests reveal little or nothing about a candidate's writing ability, can readily be handled by citing research studies. The question of how testing affects teaching, though, is less easily answered. Even the strongest proponents of multiple-choice English tests and the severest critics of essay examinations agree that no concentration on multiple-choice materials can substitute for "a long and arduous apprenticeship in actual writing" (Palmer, 1966, p. 289). But will such apprenticeship be required if, as Charles Cooper (1981) fears, a preoccupation with student performance on tests will "narrow" instruction to "objectives appropriate to the tests" (p. 12)? Lee Odell (1981) adds, "In considering specific procedures for evaluating writing, we must remember that students have a right to expect their coursework to prepare them to do well in areas where they will be evaluated. And teachers have an obligation to make sure that students receive this preparation. This combination of expectation and obligation almost guarantees that evaluation will influence teaching procedures and, indeed, the writing curriculum itself." Consequently, "our procedures for evaluating writing must be consistent with our best understanding of writing and the teaching of writing" (pp. 112-113).

Direct assessment generally addresses the skills and processes that English teachers agree are the most important aspects of writing, the focal points of instruction. Ironically, indirect assessment could affect the curriculum to its own detriment by providing an accurate measure of what it appears to neglect. Breland and Jones (1982) found a "mechanical" criterion, essay length, to be one of the very best predictors of holistic grades. This standard is wholly objective; it is very easy, quick, and inexpensive to apply; and it would yield extremely reliable ratings. An essay reading is subjective, painstaking, and costly; it can take days to finish; and the results are often quite unreliable by the usual testing standards. Yet no responsible person would suggest scoring essays merely on length, even though correlation coefficients would confirm the validity of the scorer--at least for a few administrations. Often length is a byproduct of thorough development and is achieved through use of supporting materials. But it indicates development only so long as instruction emphasizes development and not length itself. If the emphasis were merely on producing length, the quality of essay writing, and hence the correlation between essay length and holistic grade, would suffer terribly.

There is clearly more justification for using multiple-choice tests than for grading solely on length, but English instructors worry that widespread use of objective tests will cause classes to drift away from the discipline of writing essays and toward a preoccupation with recognizing isolated sentence-level errors. Were that to happen, the celebrated correlations of indirect measures with direct criterion measures would be likely to drop. Perhaps indirect measures can substitute psychometrically for direct ones, but the practice of taking multiple-choice examinations cannot substitute for the practice of writing compositions. Multiple-choice tests will remain valid and useful only if objective test skills do not become the targets of instruction. An essay component in a test acknowledges the importance of writing in the curriculum and encourages cultivation of the skill that the multiple-choice test intends to measure.

Combined measures: Both essay and objective tests entail problems when used alone. Would some combination of direct and indirect measures then be desirable? From a psychometric point of view, it is sensible to provide both direct and indirect measures only if the correlation between them is modest--that is, if each exercise measures something distinct and significant. The various research studies reviewed by Stiggins (1981) "suggest that the two approaches assess at least some of the same performance factors, while at the same time each deals with some unique aspects of writing skill...each provides a slightly different kind of information regarding a student's ability to use standard written English" (pp. 1,2). Godshalk et al. (1966) unambiguously state, "The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions...in combination with objective questions" (p. 41). The substitution of a field-trial essay read once for a multiple-choice subtest in the one-hour ECT resulted in a higher multiple correlation coefficient in six out of eight cases; when the essay score was based on two readings, all eight coefficients were higher. The average difference was .022--small but statistically significant (pp. 36-37). Not all multiple-choice subtests were developed equal, though. In every case, the usage and sentence correction sections each contributed more to the multiple prediction than did essay scores based on one or two readings, and in most cases each contributed more than scores based on three or four readings (pp. 83-84).

Apparently the success of the impromptu exercise can be repeated in large-scale assessments. Myers et al. (1966) reported that 145 readers spent five days giving each of the 80,842 ECT 20-minute essays two independent holistic readings. The operational reading proved about as reliable as comparable experimental readings. The brief essay, even when scores are based on only two readings, can make a small but unique contribution to the measurement of writing skills--something "over and beyond the prediction possible with objective measures alone" (Breland and Jones, 1982, p. 27). Accordingly, direct and indirect measures of writing ability are used together in several testing programs developed and administered by ETS. "In the English Composition Test," report Breland and Jones, "it has been found that the test and score reliability are not significantly

diminished if 20 minutes of the hour were given to a direct measure using an essay question and the remaining 40 minutes were used for 70 objective questions" (p. 3). The committee of examiners was convinced that "for the purposes of the ECT [primarily college placement], the 20-minute essay offered adequate time to sample in a general way" (pp. 3-4) the higher-order skills.

As discussed above, many critics consider 20 minutes insufficient for this purpose, but often they are thinking of the direct assessment as standing alone. Short essay and multiple-choice tests can be combined for their mutual advantage. For example, J. W. French was troubled in his 1961 study (cited in Diederich, French, & Carlton) by the lack of agreement among readers, but he observed better agreement when the readers focused on the higher-order skills that he termed "ideas," "form," and "flavor." Perhaps essay scores could be more reliable if readers were trained to concentrate just on those criteria and leave the assessment of lower-order skills to multiple-choice tests, which are far more thorough and consistent gauges of these capabilities. This division of measurement labor would make psychometric sense, not only because the essay scores might be made more reliable, but also because they would provide a unique kind of information. The Breland and Jones study (1982) underscores the feasibility of this approach. Taken together, ratings on the nine discourse, or higher-order, characteristics had exactly the same correlation with the ECT holistic scores as did the PWS holistic scores. This result "suggests that quick judgments of the nine discourse characteristics are comparable to a holistic judgment" (p. 13). In fact, multiple regression analysis showed only five of the nine discourse characteristics to be significant contributors to the prediction of ECT holistic scores. In order of importance, these were overall organization, noteworthy ideas, sentence logic, supporting material, and paragraphing and transition (p. 14).

The reliability of the direct measure can also be enhanced by combining the essay score and the objective score in one total score for reporting. Several College Board tests employ this approach. And "assessments like the ECT combine direct and indirect assessment information because the direct assessment information is not considered reliable enough for reporting. A single ECT score is reported by weighting the direct and indirect information in accordance with the amount of testing time associated with each. Such a weighting may not be the most appropriate," Breland and Jones (1982) maintain. "Another approach to combining scores might be to weight scores by relative reliabilities" (or, perhaps, validities). They add that "reliable diagnostic subscores could possibly be generated by combining part scores from indirect assessments with analytic scores from direct assessments" (p. 28).

Direct and Indirect Assessment Costs

With sufficient time and resources, many sophisticated assessment options are possible. Unfortunately, adding an essay section to a multiple-choice section can be prohibitively expensive, although desirable from a psychometric or pedagogical point of view. Some large programs simply cannot afford to pay scores of readers to be trained and to give multiple

independent readings to thousands of essays (Akeju, 1972). High correlations between direct and indirect measures may make the "significant but small" contribution of the essay seem too dearly bought (see Breland and Griswold, 1981, p. 2). Godshalk et al. (1966) declare, "It is doubtful that the slight increase in validity alone can justify the increased cost. Rather, the advantage has to be assessed in terms of the model the essay provides for students and teachers" (p. 41).

Test development costs for direct assessment can include time for planning and developing the exercises, time for preparing a scoring guide, time and money for field testing the exercises and scoring procedures, and money for producing all necessary materials. Coffman (1971) advises that if the exercise is to be scored (that is, if it is to be more than a writing sample), pretesting is essential to answer such questions as "(a) Do examinees appear to understand the intent of the question, or do they appear to interpret it in ways that were not intended? (b) Is the question of appropriate difficulty for the examinees who will take it? (c) Is it possible to grade the answers reliably" (p. 288)? One could add, (d) Will the topic allow for sufficient variety in the quality of responses? and (e) Is the time limit sufficient? Pretesting of essay topics, although extremely important, is also difficult and expensive. It "involves asking students to take time to write answers and employing teachers to rate what has been written; few essay questions are subjected to the extensive pretesting and statistical analysis commonly applied to objective test questions" (Coffman, 1971, p. 298). And yet, not even old hands at developing topics and scoring essays can reliably foresee which topics will work and which will not. Experience at ETS has shown Conlan (1982) that, after new topics have been evaluated and revised by teachers and examiners, about 1 in 10 will survive pretesting to become operational (p. 11). If the operational form includes several topics in the interest of validity, the test development costs alone will be quite high.

And yet for most direct assessments, scoring costs exceed test development costs (Stiggins, 1981, p. 5). Sometimes complex scoring criteria and procedures must be planned and tried; paid readers must be selected, recruited, and trained (not to mention housed and fed); essays must be scored; scores must be evaluated for reliability--usually at several points during the reading to anticipate problems; and results must be processed and reported. Scoring costs are related to time, of course, and the amount of time an examination reading takes will depend on the number and length of essays as well as on the number and quality of readers. It is estimated that a good, experienced holistic reader can score about 150 20- to 30-minute essays in a day. Other types of reading take longer.

Clearly, no large-scale program can afford a direct assessment that is capable of standing alone as a highly valid and reliable measure of writing ability. Such an assessment would consist of several essays in each of several modes; the essays would be written on different occasions, given at least two independent readings, and probably typed to eliminate the correlated errors arising from the papers' appearance (this procedure cost more than 50 cents per 400 word essay when McColly attempted it in

1970). Certainly it would be impossible to have each reader contribute to the score of each student, a design feature that helps explain the high reading and score reliabilities reported by Godshalk et al. (1966). And it would be "prohibitively expensive to obtain, on a routine basis, reliable subscores for such aspects of English composition as ideas, organization, style, mechanics, and creativity" (Coffman, 1971, p. 296).

For indirect assessments, test development costs typically exceed scoring costs. Many items are needed per exam, and more data are collected and analyzed from pretesting than is the case with direct assessment. Also, operational forms must be normed or equated to previous forms. Scoring is often done by computer or scanning machine, and computers can also be used to process statistical data for scaling and equating the scores. Because the collecting and scoring of multiple-choice tests is routine and relatively cheap, whereas a set of readers must be assembled and trained for each essay reading, indirect assessment allows more flexibility in planning the number, location, and time of test administrations. If test security is not breached or disclosure required, multiple-choice items can be reused so as to minimize the development costs of indirect assessment. It is probably not a good idea to reuse essay topics, which candidates can easily remember and describe to other candidates. On the other hand, disclosure will cause a larger precentage increase in cost for indirect assessment than for direct assessment, where scoring is the chief expense.

The test development cost of a 25-item operational sentence correction section in the Graduate Management Admission Test (GMAT) is about $5,840; the total cost of administering the section is between $16,000 and $17,000 (a figure that includes the expenses of editing, production, statistical analysis of pretest results, registration of candidates, transportation of materials, test administration, and the like). Four operational forms are administered each year, and so the annual cost to the GMAT program for the indirect assessment of writing ability is approximately $66,000 plus the cost of scoring and reporting. It is estimated that the total annual operating costs of including an hour-long essay examination in the GMAT would be about 20 times greater. These financial estimates illustrate Coffman's (1971) assertion that "the essay examination undoubtedly is inefficient in comparison with objective tests in terms of testing time and scoring costs" (p. 276). Even staunch advocates of direct assessment say that if an administrator "wishes to section large numbers of students according to their predicted success in composition, he probably will find a relatively sound objective test the most efficient instrument.... The expense and effort involved in reliable rating of actual compositions are not often justified where only a rough sorting of large numbers is needed" (Braddock et al., 1963, p. 44).

Compromise Measures

From the psychometric and practical, if not pedagogical, viewpoint, essay examinations may return a little extra information for a lot of money. From a pedagogical point of view, they are essential as statements

of what is important in the curriculum. An unscored writing sample offers one sort of compromise among viewpoints. Since scoring is the most expensive feature of direct assessment, overall costs could be greatly reduced by supplying test users with an unrated sample of student writing produced under standardized conditions. The unscored writing sample poses problems, however: it affords no means of ensuring that the students' performance will be assessed accurately, fairly, and consistently, or that data derived from the assessment will be used appropriately. At the very least, developers of writing samples must field test topics to make sure that they will be comparable in difficulty and satisfactory in their power to discriminate among writers at various levels of ability. In "The Story Behind the First Writing Sample," Godshalk (1961) describes the methods used to select the topics that the College Board offered in December 1960 and January 1961 for its first one-hour writing sample. The Board later discontinued the unscored exercise.

The Law School Admission Test (LSAT) has recently introduced an unscored writing sample. In the fall of 1982, all Law School Admission Council (LSAC) member law schools were surveyed regarding their use of the writing sample. Results of the questionnaire, to which 122 (64 percent) of the schools responded, show that the schools evaluate the samples and apply the information differently. A number of schools have not decided on how, or whether, to use the samples for admission decisions, but nearly all indicated that the samples are placed in applicant files. Many reported that no formal guidelines are provided to file readers. Fourteen schools evaluate the samples as part of the admission process. In most of these cases, the admission officer alone rates the papers; three schools have papers scored by legal writing instructors, one by a graduate student in English, and one by a law professor and an English professor working together. In some cases the papers are scored holistically and in others analytically (see the section on scoring methods that follows). Some of the 14 schools evaluate the writing samples of all applicants and others evaluate only the samples of "discretionary" or waiting-list applicants. Obviously, the procedures are far from uniform. Also, as should be clear from earlier discussion, the reading and score reliabilities will be low, especially across institutions. An applicant whose writing sample is sent to each of the 14 schools should expect that his or her product will receive a range of scores and serve a number of different purposes. Ninety-two schools reported that they would not formally evaluate writing samples; most will place them in files to be used as the file reader sees fit.

Another compromise measure is to use a locally developed and scored essay in conjunction with a standardized, large-scale indirect assessment. Robert Lyons (1982) and Marie Lederman (1982) recount the process by which an essay test was created and has been maintained as the sole university-wide writing assessment at CUNY (pp. 7-8, 16). A local program can fashion an essay test to its own needs and also manage scoring so as to reduce some of the costs of a large reading. Such programs, though, should be conducted with professional supervision, for they typically lack

the psychometric expertise, the research capabilities, the methodology, the quality control system, and the rigorous set of uniform procedures necessary to produce tests that are comparable in validity and reliability to those developed and scored by a major testing organization. Local programs will probably require professional assistance and training of readers to achieve sound results.

Scoring Methods for Direct Assessment

Direct assessment will, at the very least, be used to establish criterion measures of writing ability against which less expensive indirect measures are to be validated. Whatever the function of the direct assessment, several practical and psychometric issues bear on the question of how to score the essays.

Description of the methods and their uses: The three most common methods are generally called "analytic," "holistic," and "primary trait." Analytic scoring is a trait-by-trait analysis of a paper's merit. The traits--sometimes four, five, or six--are considered important to any piece of writing, regardless of context, topic, or assignment. Readers often score the traits one at a time in an attempt to keep impressions of one trait from influencing the rating of another. Consequently, a reader gives each essay he or she scores as many readings as there are traits to be rated. Subscores for each trait are summed to produce a total score. Because it is time-consuming, costly, and tedious, the analytic method is the least popular of the three. Nonetheless, it may have the most instructional value because it is meant to provide diagnostic information about specific areas of strength and weakness. The analytic method is to be preferred for most minimum competency, criterion-referenced tests of writing ability, which compare each student's performance to some preestablished standard, because the criterion (or criteria) can be explicitly detailed in the scoring guide.

A reader scoring a paper holistically as opposed to analytically reads quickly for a "global" or "whole" impression. Although specific characteristics of the essay may influence the reader, none is considered directly or in isolation. Rather than deliberate extensively, the reader assigns a comprehensive score based on his or her initial sense of the paper's overall merit. The rater does not correct, revise, or mark the paper in any way. A scoring guide and prerated sample papers called "range finders" can be used to help the reader place each paper on a rating scale. Most common are 4 and 6-point scales; generally, it is advisable to use a scale with an even number of points so that readers must score every paper as being above or below the median point of the scale. In the early days of holistic scoring, Godshalk et al. (1966) used a 3-point scale to rate criterion essays for their study. They later determined that a 4-point scale yielded higher reliabilities by relieving the congestion at the midpoint of the odd-numbered scale. In fact, the gain achieved from adding a point to the scale was roughly comparable to the gain achieved by giving the essays an additional 3-point rating (p.34). And the 4-point ratings were completed somewhat more quickly.

not clear whether 6-point or even more extended scales produce more reliable ratings than does a 4-point scale, or whether they significantly lengthen reading time.

Because it is quicker, easier, cheaper, and perhaps more reliable than the analytic method, the holistic method is more popular--in fact it is the most popular of all scoring systems. The purpose of a conventional holistic reading is to rank order a set of papers according to general quality. Hence it is to be preferred in norm-referenced tests that compare candidates to one another for the purpose of selection. Because it does not include separate scores for performance in specific areas, it has much less diagnostic and pedagogical value than the analytic method. And if the global measure says nothing about mastery or nonmastery of specific skills, the holistic method is inappropriate for criterion-referenced minimum-competency tests with cutoff scores. Such tests do not require--in fact, may work better without--a normal distribution of scores throughout the possible range.

The holistic scoring method can be modified to make it more useful for minimum-competency tests that rely on rank ordering of candidates, as do the New Jersey College Basic Skills Placement Test and the California English Placement Test. Commonly, trained readers analyze papers on the topic for attributes that characterize each level of writing skill represented by the score scale. Through group discussion, they inductively derive a scoring guide, which specifies the level of proficiency that must be demonstrated on each criterion, or writing characteristic, in the guide if an essay is to earn a given holistic point score. During the reading session, they can refer to papers that exemplify various point scores.

Readers may decide after the scoring where to set the cut score, or they may build that decision into the scoring guide. The latter procedure is more true to the spirit of a criterion-referenced minimum-competency test in that candidates are judged more according to a preestablished standard than by relative position in the group, but then a large proportion of essays could cluster around a cut score set near the midpoint of the scale. This circumstance threatens the validity and reliability with which many students are classified as masters or nonmasters because the difference of a point or two on the score total (a sum of independent readings) could mean the difference between passing and failing for the students near the cut score. As discussed, it is difficult to get essay total scores that are accurate and consistent within one or two points on the total score scale. The scoring leaders and the test user can decide to move the cut score up or down, depending on whether they deem it less harmful to classify masters as nonmasters or nonmasters as masters. Or, scorers could provide an additional reading to determine the status of papers near the cut score. This modified holistic procedure represents a sometimes difficult compromise between the need for reliable mastery or nonmastery decisions on specific criteria and the desire to permit a full response to the overall quality of an essay.

Primary trait scoring can be viewed as another sort of compromise solution, and as such it has been called a variant, or a focusing, of both the analytic and the holistic methods. Papers are scored entirely on the basis of the one trait that is deemed most important to the type of writing being assessed. For example, the primary trait could be clarity if the task is to write instructions on how to change a tire or persuasiveness if the task is to write a letter to the editor urging a position on a local issue. As Richard Lloyd-Jones (1977) explains it, primary trait scoring--unlike holistic scoring--"is not concerned with qualities of writing--syntactic fluency, for example--that are presumed to be characteristic of all good writing. Rather, it asks readers to determine whether a piece of writing has certain characteristics or primary traits that are crucial to success with a given rhetorical task" (p. 32). Primary trait scoring is more commonly used than analytic and less commonly used than holistic scoring. It is best suited for criterion-referenced testing, especially of some higher-order skill.

Advantages and disadvantages of the methods: Just as each method has preferred uses, so each has particular merits and demerits. The analytic score alone gives an account of how it was derived. Also, the method reduces "the possibility of holistic bias" (Davis et al, 1981, p. 90). A reader's overall impression of an essay may equal the strongest, or last, or most lingering impression, and a supposedly global score could reflect not general merit but the reader's overreaction to some single feature, such as an inappropriate colloquialism or a clever use of figurative language that breaks the monotony of a long reading session. On the other hand, "with analytic scoring there is a danger of numerically weighting one trait or another in a way that may not accurately reflect the relative role of that trait in the overall effectiveness of the paper" (Davis et al., 1981, p. 90). Because the scoring criteria and weights for each of the several traits are specified and set, analytic scales "are not sensitive to the variations in purpose, speaker role, and conception of audience which can occur in pieces written in the same mode" (Cooper, 1977, p. 14). After scoring papers analytically, readers frequently discover that the best or most noteworthy essay did not achieve the highest place in the rank ordering. The analytic method failed to reflect the essential and distinctive value of the composition. The analytic scoring method suggests that writing is a composite of discrete, separable aspects and that writing mastery consists of learning to perform each of the independent subtasks adequately. Readers find the method unrewarding because their judgments on the artificially isolated criteria do not sum to produce an estimate of the true overall quality of a piece of writing.

Readers also find the method slow and laborious. An analytic reading can take four, five, or six times longer than a holistic reading of the same paper (Quellmalz, 1982, p. 10). In large reading sessions, costs become prohibitive and boredom extreme. There is less time for breaks, norming sessions, and multiple independent readings, so standards may become inconsistent across readers and for the same reader over time. It is time-consuming and costly just to develop an analytic criteria

scale because, ideally, each scale is "derived inductively from pieces of writing in the mode [or exercise] for which the scale is being constructed" (Cooper, 1977, p. 14). Raters must gather, read, and discuss writing samples of varying quality in that mode in order to thrash out a scale for each aspect of composition to be rated. The analytic scoring scale that emerges from this endeavor will be complex and therefore hard to master or apply consistently. Readers employing a 4-point holistic scale will, after thorough training, have comparatively little trouble keeping track of what characterizes a paper at each level. But an analytic scoring will require readers to assimilate and keep distinct in their minds about four or five different 4-point scales (one for each trait) that they must apply sequentially. The subscores will lose validity and reliability if the distinction between judgments on different traits becomes imprecise.

Raters may more or less unconsciously assign a paper a total number of points on the basis of general impression and then apportion these points among the various "analytic" categories (Braddock et al., 1963, p. 14). As Ellis Page observes, "A constant danger in multi-trait ratings is that they may reflect little more than some general halo effect, and that the presumed differential traits will really not be meaningful" (cited in McColly, 1970, p. 151). The danger is not always realized; research by Page and by Quellmalz et al. (1982) shows that individual multitrait ratings can produce unique variance. On the other hand, research by Sarah Freedman (1979) demonstrates the powerful associative, or halo, effect that can undermine the validity of analytic subscores. Chi-square analysis revealed significant correlations of "content" with "organization" and of "sentence structure" with "mechanics" in a group of essays. The original papers had been rewritten so that an essay strong in one of the correlated traits would be weak in the other.

The holistic scoring method attempts to turn the vice of the analytic method into a virtue by using the halo effect. Coffman (1971) provides a theoretical justification for this practice: "To the extent that a unique communication has been created, the elements are related to the whole in a fashion that makes a high interrelationship of the parts inevitable. The evaluation of the part cannot be made apart from its relationship to the whole.... This [the holistic] procedure rests on the proposition that the halo effect... may actually be reflecting a vital quality of the paper" (p. 293). Lloyd-Jones (1977) adds, "One need not assume that the whole is more than the sum of the parts--although I do--for it may be simply that the categorizable parts are too numerous and too complexly related to permit a valid report" (p. 36).

The validity of analytic subscores is often suspect, and there is evidence--still inconclusive--that analytic total scores may be less reliable than holistic scores (Davis et al., 1981, p. 9), especially when a number of holistic readings are given to a paper in the time it takes to arrive at an analytic total. Coffman and Kurfman (1968) prefer "scores based on the sum of readings of several readers" to "compounding error by having a single reader assign three [or more] different scores"

(p. 106). Godshalk et al. (1966) concur: "The reading time is longer for analytical reading and longer essays, but the points of reliability per unit of reading time are greater for short topics read holistically" (p. 40).

For psychometric as well as practical reasons, then, the holistic method is usually favored for large readings. The Gary, Indiana, school system instituted a minimum-competency testing program that included a writing exercise. Essays were first scored holistically; those not meeting minimum standards were scored analytically so that the students would know what areas of their writing needed improvement. The analytic scoring proved so taxing and time-consuming that the Gary teachers replaced it with a multiple-choice test to assess performance on those aspects of writing addressed by the analytic criteria. Now all the writing samples in the testing program are scored holistically (Conlan, 1980, pp. 26-27). And perhaps all the students should take the objective test in conjunction with the essay test. As discussed above, the conventional holistic method should be modified for use in minimum-competency, criterion-referenced tests so as to measure performance against specific criteria. Still, holistic readings tend to emphasize higher-order skills (Breland & Jones, 1982). Consequently, many students above the cut score may organize and develop a topic well but need help on sentence-level skills, such as use of modifiers. Holistic scoring often regards such criteria somewhat neutrally, except in extreme cases (Breland & Jones, 1982, p. 23).

Primary trait scoring seeks to avoid such problems by focusing "the rater's attention on just those features of a piece which are relevant to the kind of discourse it is" or to the particular element of writing competency at issue. "Furthermore, a unique quality of Primary Trait scoring is that scoring guides are constructed for a particular writing task set in a full rhetorical context" (Cooper, 1977, p. 11). The holistic method, claims Richard Lloyd-Jones (1977), assumes "that excellence in one sample of one mode of writing predicts excellence in other modes--that is, good writing is good writing.... In contrast, the Primary Trait System... assumes that the writer of a good technical report may not be able to produce an excellent persuasive letter to a city council. A precise description or census of writing skills is far richer in information if the observations are categorized according to the purpose of the prose. The goal of Primary Trait Scoring is to define precisely what segment of discourse will be evaluated...and to train readers to render judgments accordingly" (p. 37).

The "richer information," though, is purchased at a cost. In sharply limiting the writing task and the range of suitable responses, the examiners make the topic less universally accessible. A situation that stimulates one writer will bore another, baffle another, and threaten still another. The results of a primary trait exercise that is not very carefully designed "will reflect experience in situations as much as skill in manipulating language forms" (Lloyd-Jones, 1977, p. 42). A primary trait writing exercise, then, would be grossly inappropriate for a population with diverse educational backgrounds and goals, but a selection of primary trait exercises, each adapted to the needs of a particular area of study, could afford an excellent means of assessing the writing skills most valued in those areas.

Again, practical issues arise. Because a writer's success depends on his or her performing in accordance with the demands of a narrowly defined situation, pretesting of primary trait exercises is especially important for determining whether the candidates will understand the situation as intended and provide an adequate range of appropriate responses. If such is not the case, the scoring guide will be unworkable. Creating a primary trait scoring guide, says Lloyd-Jones(1977), "requires patient labor, frequent trial readings, and substantial theoretical background--on the average, 60 to 80 hours of professional time per exercise, not counting the time required to administer the proposed exercise to get samples, nor the time required to try out the proposed guide" (pp. 44-45). Once a guide is accepted, it can properly be applied only to the writing exercise for which it was developed.

Summary and Conclusion

The study of writing assessment will continue to focus on the comparative strengths and limitations of direct and indirect measures because two essential psychometric properties--validity and reliability--are at issue. Direct assessment appears to be the more valid approach because it requires candidates to perform the behavior that is supposedly being measured. Consequently, it allows examiners to take a broader sampling of the skills included in the domain called "writing ability." But direct assessments may also contain more sources of invalid score variance than indirect assessments do. A candidate's essay score can reflect such irrelevant factors as the writer variable, the assignment variable, and even the handwriting variable. Indirect assessment, although more limited in range, offers greater control over item sampling error and the types of skills tested. Also, scores on different forms of standardized objective tests can be normed and equated with great precision; scores on different forms of essay tests cannot. Moreover, scores from direct assessments are typically much less reliable than scores from indirect assessments. The reliability of essay ratings improves, to the point of diminishing returns and increasing financial burden, with the number of writing exercises per student and the number of readings per exercise.

When highly reliable ratings are obtained, multiple-choice test scores can correlate highly with essay test scores, especially if there is correction for correlated errors in the direct measure. In fact, in many cases the correlation between direct and indirect measures appears to be limited chiefly by the lower reliability of the direct measure. Nonetheless, research tends to confirm common sense in showing that the two methods of assessment address somewhat different capabilities. Indirect assessment generally focuses on word- and sentence-level characteristics such as mechanics, diction, usage, syntax, and modification, whereas direct assessment often centers on higher-order or discourse-level characteristics such as statement of thesis, clarity, organization, development, and rhetorical strategy.

Comparison of different population subgroups also suggests that the two measures tap related but different skills, for although the correlation between direct and indirect measures remains significant for all groups, indirect measures tend to overpredict the essay writing performance of Black, Hispanic, and male candidates and to underpredict the essay writing performance of women. The degree of overprediction or underprediction seems to vary with ability level in different ways for different groups, but more research is needed on this point. Also, more research is needed on how the validities of proven indirect measures such as the usage and sentence correction item types would be affected by varying the mode of the direct criterion measure to make it more relevant to specific fields of study. Substantial fluctuations would show that these indirect measures are not equally useful for diverse academic disciplines. In that case, different areas of study should employ different kinds or combinations of direct and indirect measures.

Our current information, however, indicates that a program such as the GRE, which seeks to rank order applicants for selection from a large and diverse population, should use an objective test to measure writing ability. Objective tests can be highly valid, extremely reliable, easily managed, quickly scored, cost-effective, and fair to minority candidates. And as Godshalk et al. (1966) have shown, the already considerable validity of these tests can be heightened by combining the scores with verbal test scores. Where writing ability is only one of several selection criteria, a multiple-choice test can be satisfactory.

The best form of writing assessment would contain an essay as well as a multiple-choice section. Besides acknowledging the importance of essay writing in the curriculum, the addition of an essay component in a writing ability section can make a small but significant contribution to the predictive validity of the indirect measure, the extent to which the measure can indicate the quality of future writing performance. The experience of other programs with essays suggests that, given practical constraints, the GRE would be best served by a 20- or 30-minute exercise in which all candidates write on one preselected topic and all papers receive two independent holistic readings from raters who are trained to focus on higher-order skills. Unscored or locally scored writing samples are less expensive options, but also less creditable from a psychometric point of view.

The GRE Program should not consider using a short essay as the single test of writing ability; in a large-scale assessment, the essay should function as a refinement, not as the basis, of measurement for selection.

REFERENCES & BIBLIOGRAPHY

Akeju, S.A. (1972). The reliability of general certificate of education examination English composition papers in West Africa. Journal of Educational Measurement, 9, 175-180.

Atwan, R. (1982, October). Exploring the territory of multiple-choice. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, pp. 14, 19.

Bianchini, J.C. (1977). The California State Universities and Colleges English Placement Test (Educational Testing Service Statistical Report SR-77-70). Princeton: Educational Testing Service.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). Research in written composition. Urbana, IL: National Council of Teachers of English.

Breland, H.M. (1977a). A study of college English placement and the Test of Standard Written English (Research and Development Report RDR-76-77). Princeton, NJ: College Entrance Examination Board.

Breland, H.M. (1977b). Group comparisons for the Test of Standard Written English (ETS Research Bulletin RB-77-15). Princeton, NJ: Educational Testing Service.

Breland, H.M., & Gaynor, L. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

Breland, H.M., & Griswold, P.A. (1981). Group comparison for basic skills measures. New York: College Entrance Examination Board.

Breland, H.M., & Jones, R.J. (1982). Perceptions of writing skill. New York: College Entrance Examination Board.

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

Christensen, F. (1968). The Christensen rhetoric program. New York: Harper and Row.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), Educational Measurement, 2nd. ed. (pp. 271-302). Washington, DC: American Council on Education.

Coffman, W.E. (1966). On the validity of essay tests of achievement. Journal of Educational Measurement, 3, 151-156.

Coffman, W.E., & Kurfman, D.A. (1968). A comparison of two methods of reading essay examinations. American Educational Research Journal, 5, 99-107.

Conlan, G. (1980). Comparison of analytic and holistic scoring techniques. Princeton, NJ: Educational Testing Service.

Conlan, G. (1976). How the essay in the CEEB English Composition Test is scored: An introduction to the reading for readers. Princeton, NJ: Educational Testing Service.

Conlan, G. (1982, October). Planning for gold: Finding the best essay topics. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, p. 11.

Conlan, G. internal memorandum (3/24/83). Educational Testing Service.

Cooper, C.R. (Ed). (1981). The nature and measurement of competency in English. Urbana, IL: National Council of Teachers of English.

Cooper, C.R., & Odell, L. (Eds.). (1977). Evaluating writing: Describing, measuring, judging. Urbana, IL: National Council of Teachers of English.

Davis, B.G., Scriven, M., & Thomas, S. (1981). The evaluation of composition instruction. Point Reyes, CA: Edgepress.

Diederich, P.B. (1974). Measuring growth in English. Urbana, IL: National Council of Teachers of English.

Diederich, P.B., French, J.W., & Carlton, S.T. (1961). Factors in the judgments of writing ability (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service.

Dunbar, C.R., Minnick, L., & Oleson, S. (1978). Validation of the English Placement Test at CSUH (Student Services Report No. 2-78/79). Hayward: California State University.

Eley, E.G. (1955). Should the General Composition Test be continued? The test satisfies an educational need. College Board Review, 25, 9-13.

Folloman, J.C., & Anderson, J.A. (1967). An investigation of the reliability of five procedures for grading English themes. Research in Teaching of English, 190-200.

Fowles, M.E. (1978). Basic skills assessment. Manual for scoring the writing sample. Princeton, NJ: Educational Testing Service.

Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.

French, J.W. (1961). Schools of thought in judging excellence of english themes. Proceedings of the 1961 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Godshalk, F. (Winter 1961). The story behind the first writing sample. College Board Review, no. 43, 21-23.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966) The measurement of writing ability. New York: College Entrance Examination Board.

Hall, J.K. (1981). Evaluating and improving written expression. Boston: Allyn and Bacon, Inc.

Hartnett, C.G. (1978). Measuring writing skills. Urbana, IL: National Council of Teachers of English.

Huddleston, E.M. (1954). Measurement of writing ability at the college entrance level: Objective vs. subjective testing techniques. Journal of Experimental Education, 22, 165-213.

Humphreys, L. (1973). Implications of group differences for test interpretation. Proceedings of 1972 Invitational Conference on Testing Problems (pp. 56-71). Princeton, NJ: Educational Testing Service.

Huntley, R.M., Schmeiser, C.B., & Stiggins, R.J. (1979). The assessment of rhetorical proficiency: The roles of objective tests and writing samples. Paper presented at the National Council of Measurement in Education, San Francisco, CA.

Keech, C., Blickhan, K., Camp, C., Myers, M., et al. (1982). National writing project guide to holistic assessment of student writing performance. Berkeley: Bay Area Writing Project.

Lederman, M.J. (1982, October). Essay testing at CUNY: The road taken. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, pp. 7-8.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper and L. Odell (Eds.), Evaluating writing: Describing, measuring, judging (pp. 33-66). Urbana, IL: National Council of Teachers in English.

Lloyd-Jones, R. (1982, October). Skepticism about test scores. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, pp. 3, 9.

Lyman, H.B. (1979). Test scores and what they mean. Englewood Cliffs, NJ: Prentice-Hall.

Lyons, R.B. (1982, October). The prospects and pitfalls of a university-wide testing program. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, pp. 7, 16.

McColly, W. (1970). What does educational research say about the judging of writing ability? The Journal of Educational Research, 64, 148-156.

Markham, L.R. (1976). Influences of handwriting quality on teacher evaluation of written work. American Educational Research Journal, 13, 277-283.

Michael, W.B., & Shaffer, P. (1978). The comparative validity of the California State Universities and Colleges English Placement Test (CSUC-EPT) in the prediction of fall semester grade point average and English course grades of first-semester entering freshmen, Educational and Psychological Measurement, 38, 985-1001.

Moffett, J. (1981). Active voice: A writing program across the curriculum. Montclair, NJ: Boynton/Cook.

Mullis, I. (1979). Using the primary trait system for evaluating writing. National Assessment of Educational Progress.

Myers, A., McConville, C., & Coffman, W.E. (1966). Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 26, 41-54.

Myers, M. (1980). Procedure for writing assessment and holistic scoring. Urbana, IL: National Council of Teachers in English.

Notes from the National Testing Network in Writing. (1982, October) New York: Instructional Resource Center of CUNY.

Noyes, E.S. (Winter 1963). Essays and objective tests in English. College Board Review, no. 49, 7-10.

Noyes, E.S., Sale, W.M., & Stalnaker, J.M. (1945). Report on the first six tests in English composition. New York: College Entrance Examination Board.

Odell, L. (1981). Defining and assessing competence in writing. In C.R. Cooper (Ed.), The nature and measurement of competency in English (pp. 95-138). Urbana, IL: National Council of Teachers of English.

Odell, L. (1982, October). New questions for evaluators of writing. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, pp. 10, 13.

Palmer, O. (1966). Sense or nonsense? The objective testing of English composition. In C.I. Chase & H.G. Ludlow, (Eds.). Readings in Educational and Psychological Measurement (pp. 284-291). Palo Alto: Houghton.

Purves, A., et al. (1975). Common sense and testing in english. Urbana, IL: National Council of Teachers in English.

Quellmalz, E.S. (1980), November). Controlling rater drift. Report to the National Institute of Education.

Quellmalz, E.S. (1982). Designing writing assessments: Balancing fairness, utility, and cost. Los Angeles: UCLA Center for the Study of Evaluation.

Quellmalz, E.S., Capell, F.J., & Chih-Ping, C. (1982). Effects of discourse and response mode on the measurement of writing competence. Journal of Educational Measurement, 19, 241-258.

Quellmalz, E.S., Spooner-Smith, L.S., Winters, L., & Baker, E. (1980, April). Characteristics of student writing competence: An investigation of alternative scoring schemes. Paper presented at the National Council of Measurement in Education.

Spandel, V. & Stiggins, R.J. (1981). Direct measures of writing skills: Issues and applications. Portland, OR: Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory.

Stiggins, R.J. (1981). An analysis of direct and indirect writing assessment procedures. Portland, OR: Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory.

Stiggins, R.J. & Bridgeford, N.J. (1982). An analysis of published tests of writing proficiency. Portland, OR: Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory.

Thompson, R.F. (1976). Predicting writing quality. English Studies Collections, 1 no. 7. East Meadow, NY.

Tufte, V. & Stewart, G. (1971). Grammar as style. New York: Holt, Rinehart and Winston, Inc.

Werts, C.E., Breland, H.M., Grandy, J., & Rock, D. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. Educational and Psychological Measurement, 40, 19-29.

White, E.M. (1982, October). Some issues in the testing of writing. Notes from the National Testing Network in Writing. New York: Instructional Resource Center of CUNY, p. 17.

Willig, M.H. (1926). Individual diagnosis in written composition. Journal of Educational Research, 13, 77-89.

Winters, L. (1978, November). The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education.

Wiseman, S., & Wrigley, S. (1958). Essay-reliability: The effect on choice of essay-title. Educational and Psychological Measurement, 18, 129-138.

5402003 • L54P1.5 • 297922 • Printed in U.S.A.